The influence of reward magnitude on stimulus memory and stimulus generalization in categorization decisions

Schlegelmilch, René ; von Helversen, Bettina

Abstract: Reward magnitude is a central concept in most theories of preferential decision making and learning. However, it is unknown whether variable rewards also influence cognitive processes when learning how to make accurate decisions (e.g., sorting healthy and unhealthy food differing in appeal). To test this, we conducted 3 studies. Participants learned to classify objects with 3 feature dimensions into two categories before solving a transfer task with novel objects. During learning, we rewarded all correct decisions, but specific category exemplars yielded a 10 times higher reward (high vs. low). Counterintuitively, categorization performance did not increase for high-reward stimuli, compared with an equal-reward baseline condition. Instead, performance decreased reliably for low-reward stimuli. To analyze the influence of reward magnitude on category generalization, we implemented an exemplar-categorization model and a cue-weighting model using a Bayesian modeling approach. We tested whether reward magnitude affects (a) the availability of exemplars in memory, (b) their psychological similarity to the stimulus, or (c) attention to stimulus features. In all studies, the evidence favored the hypothesis that reward magnitude affects the similarity gradients of high-reward exemplars compared with the equal-reward baseline. The results from additional reward-judgment tasks (Studies 2 and 3) strongly suggest that the cognitive processes of reward-value generalization parallel those of category generalization. Overall, the studies provide insights highlighting the need for integrating reward- and category-learning theories.

DOI: https://doi.org/10.1037/xge0000747

The Influence of Reward Magnitude on Stimulus Memory and Stimulus Generalization in Categorization Decisions

René Schlegelmilch[1,2] and Bettina von Helversen[1,2]

[1]University of Bremen

[2]University of Zurich

https://doi.org/10.1037/xge0000747

Contacts:

René Schlegelmilch: r.schlegelmilch@uni-bremen.de

Bettina von Helversen: b.helversen@uni-bremen.de

Correspondence concerning this article should be addressed to René Schlegelmilch, University of Bremen, Department of Psychology, COGNIUM, Hochschulring 18, 28359 Bremen, Germany. E-mail: r.schlegelmilch@uni-bremen.de

All data files are available on the OSF at DOI 10.17605/OSF.IO/GTJV3. Parts of this research were presented at the annual meeting of the SJDM (Vancouver), the Annual Meeting of the Psychonomic Society (Vancouver), the TeaP (Dresden and Marburg), and the JDMx Meeting for Early-Career Researchers (Bonn and Konstanz).

Abstract

Reward magnitude is a central concept in most theories of preferential decision making and learning. However, it is unknown whether variable rewards also influence cognitive processes when learning how to make accurate decisions (e.g., sorting healthy and unhealthy food differing in appeal). To test this, we conducted three studies. Participants learned to classify objects with three feature dimensions into two categories before solving a transfer task with novel objects. During learning, we rewarded correct decisions, but specific category exemplars yielded a 10 times higher reward (high vs. low). Counterintuitively, categorization performance did not increase for high-reward stimuli, compared to an equal-reward baseline condition. Instead, performance decreased reliably for low-reward stimuli. To analyze the influence of reward magnitude on category generalization, we implemented an exemplar-categorization model and a cue-weighting model using a Bayesian modeling approach. We tested whether reward magnitude affects (a) the availability of exemplars in memory, (b) their psychological similarity to the stimulus, or (c) attention to stimulus features. In all studies, the evidence favored the hypothesis that reward magnitude affects the similarity gradients of high-reward exemplars compared to the equal-reward baseline. The results from additional reward-judgment tasks (Studies 2 and 3) strongly suggest that the cognitive processes of reward-value generalization parallel those of category generalization. Overall, the studies provide insights highlighting the need for integrating reward- and category-learning theories.

Keywords: Category Learning, Reward Learning, Generalization, Similarity, Memory, Decision Making

The Influence of Reward Magnitude on Stimulus Memory and Stimulus Generalization in Categorization Decisions

Since Thorndike (1911) formulated the law of effect, decades of reinforcement learning research (see Mackintosh, 1974; Sutton & Barto, 1998) have affirmed that knowing the reward values of objects or situations is vital for optimizing decision-making success (e.g., Kahnt, 2018; Tobler, Fiorillo, & Schultz, 2005), with success usually defined as obtaining the largest rewards (e.g., by predicting reward values; see also Glimcher, Camerer, Fehr, & Poldrack, 2009; Kahnt, 2018; Schultz, 2006; Schultz & Dickinson, 2000; for related overviews). However, while the motivating or reinforcing role of reward in "successful" decisions has been widely studied (see D. Lee, Seo, & Jung, 2012; Levine & Edelstein, 2009; O'Doherty, Cockburn, & Pauli, 2017; Schultz, 2006), how reward magnitude relates to the *accuracy* of decision making has been much less studied (see Jenkins, Mitra, Gupta, & Shaw, 1998; Seger & Peterson, 2013).

In many professional domains, as well as in everyday life, reward often depends on decision accuracy. For instance, physicians differentiate between types of cancer to accurately assign treatments, and animal caretakers differentiate between types of food for different animals. In these decisions, receiving a reward for the decision (i.e., recovering patients or healthy animals) usually serves as feedback about decision accuracy, which allows learning about how stimuli relate to categories (e.g., the food digestible by one animal species but not by another). However, so far, it has been neglected that reward magnitudes may vary between specific category instances independent of how stimulus categories are objectively defined (e.g., more or less enjoyable food). In light of studies showing how easily variable rewards distract from solving cognitive tasks (e.g., B. A. Anderson, 2013; Della Libera & Chelazzi, 2009), the question arises whether reward magnitude also influences category learning and thereby decision accuracy.

The goal of the current research was to investigate whether and how stimulus-specific reward magnitude affects performance in category learning (e.g., learning speed and accuracy) and category generalization (i.e., category inference for novel stimuli) as well as to identify the involved cognitive mechanisms. These research questions, however, concern not only category learning but also several decision

domains, since various aspects of categorization decisions are shared by perceptual, social, and economic decisions, such as the memorization and identification of objects, probability learning, inferential choice, and multi-attribute integration (e.g., Achtziger, Alós-Ferrer, Hügelschäfer, & Steinhauser, 2015; Goldstein, 1991; Juslin, Jones, Olsson, & Winman, 2003; Markman & Ross, 2003; Miendlarzewska, Bavelier, & Schwartz, 2016; Pachur & Olsson, 2012; Russell et al., 1999; Seger, Braunlich, Wehe, & Liu, 2015; Seger & Peterson, 2013; E. R. Smith & Zarate, 1992).

Typical category-learning research implements reward in terms of corrective feedback (i.e., present = correct vs. absent = incorrect, also referred to as supervised learning; e.g., Sutton & Barto, 1998). In the following three sections, we outline three ways that variations in reward value could influence category learning beyond feedback (for further discussions, see Seger & Peterson, 2013; Tricomi & Fiez, 2008). We explain the resulting predictions concerning changes in categorization accuracy from three theoretical perspectives. We first view category inference from an exemplar memory perspective (e.g., Medin & Schaffer, 1978; Nosofsky, 1986) with an emphasis on stimulus recall, then turn to the similarity-based generalization of stimulus categories and reward values (e.g., Kahnt, Park, Burke, & Tobler, 2012; Shepard, 1987; Wimmer, Daw, & Shohamy, 2012), and finally define how reward value might alter attention paid to stimulus dimensions, including consideration of rule-based categorization strategies (e.g., B. A. Anderson, 2013; Juslin, Jones, et al., 2003; Le Pelley, Mitchell, Beesley, George, & Wills, 2016; Reed, 1972).

Reward and Exemplar Memory Strength

One way to learn stimulus categories (or their reward) is to store single instances in memory, as assumed in the popular generalized context model (GCM; Medin & Schaffer, 1978; Nosofsky, 1986, 1988b; see Modeling section for details). According to the GCM, a decision maker recalls previously encountered exemplars and compares them to the current stimulus. The stimulus is categorized according to its similarity to the recalled exemplars. Within this process, reward magnitude could influence which or how easily exemplars are recalled. In this vein, recent studies from the domain of working memory have shown that reward anticipation is positively related to recall accuracy (among other behavioral measures; see also Miendlarzewska et al., 2016). For instance, Wolosin, Zeithamova, and Preston (2012) found that after seeing pictures of either high- or

low-reward value, participants' better recalled the high-reward pictures (see also Aberg, Müller, & Schwartz, 2017). It seems reasonable to assume a similar effect in category learning, as working memory capacity has been shown to predict category-learning performance (Craig & Lewandowsky, 2012; Lewandowsky, 2011; Lewandowsky, Yang, Newell, & Kalish, 2012), and working memory and category learning rely on shared cognitive processes (see also Oberauer, 2009; Oberauer, Süß, Wilhelm, & Sander, 2007).

One explanation for these results is the memory-strength hypothesis, which builds on the assumption that increasing reward enhances the availability of the associated exemplars in memory (e.g., described by the memory-strength parameter in the GCM; Nosofsky, 2011). This hypothesis is similar to the idea that presenting a stimulus more frequently during learning increases the memory strength of the stimulus during a subsequent categorization task (e.g., Nosofsky, 1988b, 1991b). This formal concept has also been applied to describe the likelihood of stimulus recall in working memory tasks (e.g., Nosofsky, Little, Donkin, & Fific, 2011). Hence, the memory-strength hypothesis (implemented in the GCM) predicts an increase in categorization accuracy for high-reward items, compared to an equal-reward baseline condition. However, it also predicts a proportional decrease in accuracy for low-reward stimuli, as the GCM formally treats memory strength as a trade-off parameter (increasing memory strength for some exemplars simultaneously decreases memory strength for other exemplars; see Modeling section for details).

The memory-strength hypothesis is formally equivalent to assuming that low-reward items are remembered but strategically ignored during the decision process, which might be a plausible strategy. This view includes the possibility that high-reward exemplars become the sole representatives of their categories. In this extreme case, the GCM would become formally identical to a prototype model (e.g., Medin & Smith, 1981; Reed, 1972; J. D. Smith & Minda, 1998). However, while traditionally a prototype represents the most typical stimulus of a category, in this case the prototype would represent the most rewarded one. Thus, the predictions of the memory-strength hypothesis cover three different yet related theoretical ideas of the effect of variable rewards in category learning: (a) enhanced stimulus memory, (b) strategic information integration, and (c) reward prototypes.

Despite the common intuition about the enhancing effects of reward motivation, a

closer look at recent working memory studies (e.g., Allen & Ueno, 2018; Klink, Jeurissen, Theeuwes, Denys, & Roelfsema, 2017; Wallis, Stokes, Arnold, & Nobre, 2015) reveals that differences in performance between high- and low-reward stimuli possibly stem from worse performance for low-reward stimuli instead of better performance for high-reward stimuli. For instance, Klink et al. (2017) compared conditions in which item reward values were disclosed either during item encoding or during retrieval. Their data show that low-reward stimuli were recalled worse in the encoding condition compared to the retrieval condition, while performance for high-reward stimuli did not differ between conditions. This pattern of results indicates that reward magnitude could affect categorization performance through other cognitive processes than memory strength, which we describe below.

Reward and Stimulus Generalization

In decision making, the term generalization describes the ability to (nonrandomly) react to novel stimuli after observing the consequences for other stimuli. For instance, in a typical reward-learning task, the probability of responding (or expecting a reward) when seeing an unknown stimulus is a function of its similarity to known stimuli. One widely acknowledged theory describing this function is Shepard's (1987) law of stimulus generalization (see also Jäkel, Schölkopf, & Wichmann, 2008a; Mackintosh, 1974; Miendlarzewska et al., 2016; Seger & Peterson, 2013; Wimmer et al., 2012; Wimmer & Shohamy, 2012). It assumes that generalization declines exponentially with psychological similarity between the current stimulus and the previously trained "reward" stimulus. The speed of the decline is governed by a free parameter, the similarity gradient.

Shepard's (1987) stimulus generalization is an important part of both successful models of working memory predicting recall errors via similarity-based interference (e.g., Jonides et al., 2008; Oberauer & Lin, 2017) and context models of categorization. In the latter, the presented stimulus is categorized based on its similarity to previously encountered instances (Medin & Schaffer, 1978; Nosofsky, 1986), or category prototypes (Medin & Smith, 1981; J. D. Smith & Minda, 1998). Here the similarity gradient indicates the weight assigned to differences between a stimulus and exemplars in memory, such that category inference with narrow gradients includes mainly highly

similar exemplars (stronger discrimination), while broad gradients also include more distant exemplars (stronger integration).

Shepard (1987) suggested that "differential reinforcement could shape the generalization function and contours around a particular stimulus" (Shepard, 1987, p. 237; see also Mackintosh, 1974). This statement implies that the similarity gradients of specific instances could depend on their reward value. Indeed, the results of a recent reward-learning study by Kahnt et al. (2012) showed that reward-dependent generalization can be described by assuming narrower similarity gradients for high-reward cues than for low-reward cues (see Miendlarzewska et al., 2016, for further discussions; see also Wimmer & Shohamy, 2012). In other words, the generalization of high reward value was "narrower" for trained high-reward cues (stronger discrimination).

It seems likely that reward magnitude influences the similarity gradients of category instances, as category feedback processing largely overlaps with reward (value) processing on the neural level (Aron et al., 2004; Daniel & Pollmann, 2010; Miendlarzewska et al., 2016; Seger et al., 2015; Seger & Peterson, 2013; Seger & Spiering, 2011; Shohamy, Myers, Kalanithi, & Gluck, 2008). However, a formal transfer from reward learning to models of categorization leads to implausible predictions. Specifically, assuming a GCM with narrower similarity gradients for high-reward exemplars than for low-reward exemplars, as suggested by the data of Kahnt et al. (2012), would predict improved learning of low-reward items, which seems unlikely.

One reason for this apparent contradiction could be that the effect of specific manipulations on similarity gradients depends on the task structure. For instance, Hendrickson, Perfors, Navarro, and Ransom (2019) found that when participants learned to categorize stimuli into "Category A or not" (similar to learning "reward or not"), increasing the sample size (from 4 to 12) led to behavior compatible with the idea of narrower similarity gradients (stronger discrimination). In contrast, when participants learned about two categories ("Category A vs. B"), increasing the sample size for Category A induced a category bias, which is consistent with the assumption of broader similarity gradients for Category A exemplars (see also Goldstone & Son, 2005; Homa, Blair, McClure, Medema, & Stone, 2018; Polk, Behensky, Gonzalez, & Smith, 2002).

Moreover, in a study on reward and category preferences, Maddox and Bohil (2001) reported a very similar category bias when one of two categories was associated with a higher reward (see also Figure 2 in Healy & Kubovy, 1981), as Hendrickson et al. (2019) reported for stimulus frequency with two exclusive categories. This suggests that in a categorization task with two exclusive categories, the effect of reward on similarity gradients might also be in the opposite direction, as in the "one category" reward-learning task by (Kahnt et al., 2012); that is, the similarity gradients of high-reward items may become broader instead of narrower. This hypothesis is generally compatible with recent theories about how reward value could affect category generalization (Miendlarzewska et al., 2016; Seger & Peterson, 2013) and the idea that reward magnitude modulates *how informative* an instance of feedback is perceived to be, as also discussed by Della Libera and Chelazzi (2006) for the case of perceptual decision making.

In our studies, we focused on category structures with few high-reward stimuli, because this allowed us to distinguish between the predictions of the outlined similarity and memory-strength hypotheses. In these tasks, the similarity hypothesis predicts a decrease in performance for low-reward exemplars but unchanged performance for high-reward exemplars, compared to baseline (see Modeling section for details). In contrast, the memory-strength hypothesis predicts an increase in accuracy for high- and a decrease for low-reward exemplars. However, under different category structures, the behavioral implication of the formal assumptions might change (see also the section Boundary Conditions, Rules, and Exceptions in the General Discussion). Last, besides similarity or stimulus recall, reward magnitude could also affect which stimulus dimensions a decision maker attends to (i.e., dimension attention) while making categorizations (e.g., evaluating similarity for color, but not for size; see Le Pelley et al., 2016; Nosofsky, 1988b).

Reward, Feature Attention, and Rule-Based Strategies

One fundamental assumption of reward and reinforcement learning theories is that dimension attention (also referred to as feature or cue attention) corresponds to how well a dimension predicts rewards or decision outcomes (Sutton & Barto, 1998). Indeed, in standard category-learning tasks, participants shift attention to those stimulus

features that reliably predict category membership (e.g., Blair, Watson, & Meier, 2009; Chen, Meier, Blair, Watson, & Wood, 2013; Matsuka & Corter, 2008; Rehder & Hoffman, 2005a, 2005b). According to a recent review by Le Pelley et al. (2016), however, dimension attention also depends on the value of the predicted outcome. Thus, when introducing variable rewards in category learning, competing task goals of predicting category membership or reward magnitude could affect which stimulus dimensions are attended to.

If high-reward stimuli automatically draw more attention than low-reward stimuli (e.g., B. A. Anderson, 2013) and people pay more attention to dimensions that are predictive of category membership, it seems likely that attention would focus on the dimensions that predict the category memberships of high-reward exemplars. However, the dimensions that best predict the category membership of the high-reward stimuli might not adequately predict the categories of *all* exemplars. Thus, the attention hypothesis implies, similar to the memory-strength hypothesis, that categorization accuracy increases for high-reward stimuli at the cost of accuracy for other exemplars. In contrast to the similarity and memory-strength hypotheses that apply only to approaches that concern stimulus-specific effects, the attention hypothesis also applies to an important alternative approach to categorizations, namely, rule-based or cue-abstraction models (CAMs; see Ashby & Gott, 1988; Barsalou, 1990; Goldstein, 1991; Medin & Schaffer, 1978; Medin & Smith, 1981; Reed, 1972; for related studies).

We use the term cue abstraction (e.g., Juslin, Jones, et al., 2003; Juslin, Olsson, & Olsson, 2003) to refer to a class of models that assume people learn how object dimensions are related to the criterion of interest (e.g., longer objects are more likely in Category A). This includes very simple rule strategies (e.g., Ashby & Gott, 1988), in which the values of only one or two dimensions are used to infer categories (e.g., long = Category A, short = Category B). Such simple rules, however, can be seen as special cases of so-called linear attribute-weighting models (see also Bröder, Newell, & Platzer, 2010; Juslin, Jones, et al., 2003), which are often used to describe attitude formation in clinical, social, and quantitative judgments (see N. Anderson, 1981; Brehmer, 1976; Dawes & Corrigan, 1974; Hoffmann, von Helversen, & Rieskamp, 2014). These models assign a weight to each object dimension that reflects the models' assumptions regarding the strength of the dimension–criterion relation—and thus the relative importance of

these dimensions. To predict a category, the weighted values of the dimensions are then added (i.e., without assuming interactions) and compared to a decision criterion (Hoffmann, von Helversen, & Rieskamp, 2016; Juslin, Jones, et al., 2003).

The predictions of the attention hypothesis from the cue-abstraction perspective are analogous to those from the context model perspective, despite making different assumptions about how stimulus–response contingencies are represented. If people pay more attention to high-reward stimuli than low-reward stimuli (e.g., B. A. Anderson, 2013) they might give more weight to changes on dimensions that predict correct classifications of high-reward stimuli. This, in turn, should lead to an increase in categorization accuracy for high-reward items, compared to baseline. Moreover, if the ignored dimensions help predict the categories of low-reward stimuli, then categorization accuracy should decrease for low-reward stimuli.

Although memory- and rule-based approaches have often been contrasted, a growing number of researchers acknowledge that both might be part of a more general cognitive framework, as exemplified in hybrid and mixture models (e.g., Bröder, Gräf, & Kieslich, 2017; Erickson & Kruschke, 1998; Johansen & Palmeri, 2002; Medin, Altom, & Murphy, 1984; Nosofsky, Palmeri, & McKinley, 1994; Vanpaemel & Storms, 2008). In addition, a number of different representational or mechanistic approaches, such as category clustering (e.g., Love, Medin, & Gureckis, 2004), decision trees (e.g., Lafond, Lacouture, & Cohen, 2009), and Bayesian hypothesis testing (e.g., Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010), have been proposed. In the current research, we focused on the context model (GCM; Nosofsky, 2011) and cue-abstraction approaches (e.g., Bröder et al., 2010; Hoffmann et al., 2016; Juslin, Jones, et al., 2003) to describe the consequences of stimulus-specific reward magnitude for category inference. Specifically, in the GCM we examined whether effects of variable rewards are captured by changes in memory strengths or similarity gradients (GCM; Nosofsky, 2011) and contrasted this exemplar account against a CAM (see Modeling section).

Study Overview

In the following we present three category-learning studies in parallel, in which we investigated whether reward magnitude affects (a) instance memory strength or (b)

similarity-based category generalization from an exemplar perspective, and/or (c) the weight given to feature dimensions (feature attention) from both an exemplar and a cue-abstraction perspective. We first summarize the design, hypotheses, and planned tests across the three conducted studies, while highlighting additional exploratory questions in the course of the analyses (e.g., on the interplay between reward magnitude and how representative an instance is of its category [typical vs. atypical] in Study 3). In the Results, we analyze the effects of reward magnitude on categorization accuracy for repeatedly trained category items. In the Modeling section, we turn to the question of how reward magnitude affects category generalization (including novel items), in which we differentiate potential strategies and parameter changes in a Bayesian latent mixture modeling approach. Finally, a joint analysis of categorization decisions and behavior in reward-judgment tasks addresses the question of how category generalization might be linked to learning and generalizing the reward value of category instances.

The tests for the hypotheses were preregistered on the Open Science Framework (OSF; osf.io) for Study 1 (osf.io/rkpxd) and Study 2 (osf.io/c3zqx); we made some slight modifications, which are highlighted in the analyses. We also provide Online Supplemental Material on the OSF, including all data and analysis scripts (R code; R Development Core Team, 2008). The preregistrations also concern tests for decision times and more detailed tests on the transfer-phase behavior, which are beyond the scope of this paper but are provided in the Online Supplemental Material, as well as a study-wise presentation of the methodological details of each study.[1]

Design and Hypotheses

In all three studies, participants learned how to categorize fictitious objects into two categories, before applying their category knowledge to novel items in a transfer phase. We varied between participants whether the reward magnitude for making

[1]Please note, we do not report the studies in chronological order to improve readability as suggested by an anonymous reviewer. We first carried out Study 3, for which there is no preregistration, and then focused on replicating the main results in Studies 1 and 2, in this order. In Study 2, we also extended the design to investigate the effect of penalty magnitude, which was, however, for exploration and is not reported, also due to the lack of a valid baseline (equal penalties). The corresponding data can be found on the OSF.

correct categorizations varied between specific training exemplars (high- and low-reward exemplars), or whether a correct categorization yielded the same reward for all exemplars.

We tested the following hypotheses: First, increasing reward magnitude increases instance memory strength (i.e., exemplar-specific V in the GCM; see Nosofsky, 2011, and the Modeling section), which can be interpreted as an increased retrieval rate for high-reward exemplars, or as strategic ignorance of low-reward exemplars during category inference. An increased memory strength for trained high-reward exemplars predicts an increase in their categorization accuracy but a decline in accuracy for low-reward exemplars, compared to baseline performance. This hypothesis also covers the idea that high-reward category exemplars are perceived as "more prototypical" (see also Medin & Smith, 1981).

Second, if categories are generalized from exemplars to other (novel) stimuli, then reward magnitude could affect exemplar-specific similarity gradients. This idea is based on two assumptions: that similarity gradients can be stimulus specific (see also Goldstone, Steyvers, & Larimer, 1996; Rodrigues & Murre, 2007), and that reward magnitude influences generalizations of value (Kahnt et al., 2012). If similarity gradients of high-reward items become broader, this would predict a decrease in categorization accuracy for low-reward items, without increasing accuracy for high-reward items, compared to baseline performance.

The third hypothesis concerns the distribution of attention over stimulus features or dimensions during category inference. Dimension attention is usually assumed to correspond to how well a dimension predicts category membership (see Pothos & Wills, 2011) as well as rewards (see Le Pelley et al., 2016). With variable stimulus rewards, high-reward stimuli might draw more attention than low-reward stimuli, which could lead people to attend more strongly to those stimulus features that predict the category membership of high-reward stimuli.

In Studies 2 (preregistered) and 3 (exploratory) we additionally investigated whether the influence of reward magnitude on categorization behavior is related to the generalization of reward, that is, the expected reward for categorizing a stimulus correctly (e.g., Kahnt et al., 2012; Wimmer et al., 2012). In particular because reward learning can be described by narrower similarity gradients for high-reward items, while

our similarity hypothesis assumes broader similarity gradients for high-reward items, we wanted to investigate if reward judgments and categorizations can be described by the same modeling account. Specifically, we assumed that the exemplar similarity gradients estimated in the category-generalization task can be used to predict subsequent judgments about the reward value of trained and novel stimuli.

General Method and Materials

The experiments were programmed using javaScript (jsPsych; de Leeuw, 2015). Studies 1 and 2 were conducted online on Amazon Mechanical Turk. Study 3 was conducted in the laboratory and presented on standard desktop computers in single-seat cubicles. The study design and methods were approved by the ethics committee of the University of Zurich.

Tasks, manipulations, and stimuli. In all studies, participants learned how to feed two animal species (Tami vs. Humi) with a limited number of food items (10 in Studies 1 and 3 and 12 in Study 2), varying on three dimensions, each with quasicontinuous values. Each item belonged to only one category, and learning proceeded in single-item trials with corrective feedback after every decision (see Figure 1, Tables 1 and 2). The items were repeatedly presented in random order within 10, 10, and 12 training blocks, in Studies 1, 2, and 3, respectively. Each correct categorization yielded an immediate monetary reward in "thalers," which was collected and exchanged for a bonus payment after completion. In the control conditions (C in Tables 1 and 2) all items rendered equal rewards, serving as a baseline reference for item-specific performance changes in the reward conditions (between-subjects comparisons). In the reward condition, two (Studies 1 and 2) or four (Study 3) items rendered a 10 times higher reward (high-reward items) than the other items (low-reward items). The high-reward items were equally distributed across categories to preclude category preferences (e.g., Maddox & Bohil, 2001). In Study 3, we included two reward conditions to counterbalance to which item sets (1 vs. 2) the high rewards were assigned (see Table 2).

In general, the training-item categories were determined by a CAM (logistic response function), based on the (mean-centered and equal-weighted) sum of the stimulus dimension values, followed by a median split. Similar task structures have

been previously used to investigate category learning and are usually found to induce exemplar-based rather than cue-abstraction processes (or rule-based strategies; e.g., Ashby, Maddox, & Bohil, 2002; Bröder et al., 2010; Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Hoffmann et al., 2016; Juslin, Jones, et al., 2003). In addition, Daniel and Pollmann (2010) used a similar task when comparing neural activation patterns of feedback and reward processing. Our task, however, differed from structurally similar information-integration tasks (e.g., Ashby & Gott, 1988). For instance, we repeatedly presented stimuli and used rather discrete (though quasicontinuous) features, because one of our goals was to test differences in item identification. The visual representations of the feature values are illustrated in Figure 1. In Studies 1 and 3, the visual feature-to-dimension assignments were fixed, but they were randomly balanced in Study 2 (between participants). Similarly, in Study 3, the six possible visual arrangements of the dimensions (left, middle, right location) were randomly balanced between participants. The response button positions (see Figure 1) were randomly balanced across participants in each study, as well.

A transfer phase followed that included trained and novel items (see Appendix A, Tables A1–3), randomly presented in 6, 8, and 10 repeated transfer blocks, in Studies 1, 2, and 3, respectively. The trials were identical to the training trials but without feedback, and each ended automatically after the decision. Correct classifications of the trained items resulted in the same (covert) reward as during training. Novel items yielded 3 thalers in any case (undisclosed to the participants) because there was no objective accuracy criterion to determine their category membership.

To reduce the length of the online studies (Studies 1 and 2) we included only four of the trained items in the set of transfer stimuli. This always contained the high-reward items (i.e., Set 'High' in Table 1), and one random low-reward training item per category. Thus, trained high- versus low-reward items were balanced in number. This also gives more weight to the novel items in the cognitive model analyses in Studies 1 and 2 compared to Study 3, in which we presented all 10 training items in the transfer phase again.

We applied additional constraints to the stimulus sets to increase the strength of our hypothesis tests. First, as depicted in Tables 1 and 2, the number of high-reward items was generally low (i.e., 2, 2, and 4, in Studies 1, 2, and 3, respectively). Thus, it

Table 1

Training Stimuli and Manipulations in Studies 1 and 2

Food item	Dimension			Reward		Set
	1	2	3	C	R	
Study 1						
Humi						
1	4	3	1	3	30	High
2	2	2	4	3	3	Low
3	3	1	4	3	3	Low
4	4	1	3	3	3	Low
5	1	2	5	3	3	Low
Tami						
6	5	1	4	3	30	High
7	3	5	4	3	3	Low
8	4	1	5	3	3	Low
9	2	5	3	3	3	Low
10	5	3	2	3	3	Low
Study 2						
Humi						
1	2	4	1	3	30	High
2	4	2	2	3	3	Low
3	5	1	2	3	3	Low
4	2	1	4	3	3	Low
5	2	2	4	3	3	Low
6	1	4	2	3	3	Low
Tami						
7	4	1	5	3	30	High
8	5	4	2	3	3	Low
9	4	5	2	3	3	Low
10	2	5	4	3	3	Low
11	2	4	4	3	3	Low
12	4	2	4	3	3	Low

Note. The three dimensions were represented by the visual stimulus features (see Figure 1). Final sample sizes after exclusions (see Data Cleansing) were $N = 90$ in the Study 1 control (baseline) condition and $N = 90$ in the Study 1 reward condition, and $N = 61$ in the Study 2 control condition and $N = 80$ in the Study 2 reward condition. Set summarizes the stimulus rewards as used for indexing in the Results and Modeling sections. C = control; R = reward.

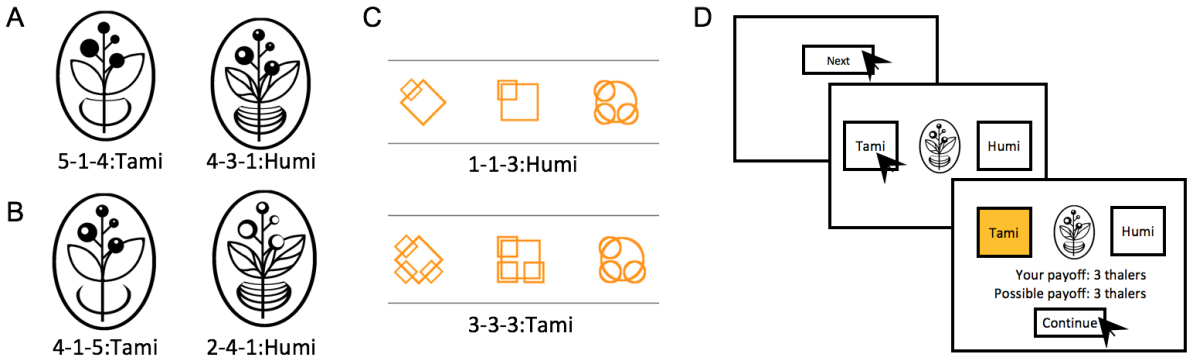


Figure 1. Example stimuli for Studies 1 (A), 2 (B), and 3 (C), and a training-trial illustration for Studies 1 and 2 (D). (A, B) High-reward items in Studies 1 and 2. Variations in the shading of the berries and number of lines in the leaves and base represent the values of the dimensions (i.e., 1, 2, and 3 in Table 1). (C) Each row represents one food item. Shape counts represent dimension values. (D) Clicking on "Next" uncovered an item. After choosing an animal to feed (Tami vs. Humi), the feedback was provided by coloring the correct category in orange. In the case of an incorrect decision, the chosen option turned blue. Feedback and reward were uncovered together with a "Continue" button. In Studies 1 and 2 (online), the potential reward was always uncovered with feedback to ensure reward encoding. In Study 3 (laboratory), the potential reward was depicted before each trial (1 s; without "Next" button), and the received reward was provided with feedback.

should have been easy to memorize the high-reward stimuli (if desired) in Studies 1 and 2. Second, in Study 3, the atypical high reward set (AR-High) also contained two items that were atypical in the given category structure (i.e., they were exceptions to the cue-abstraction predictions). High reward was associated with the atypical items in the atypical reward (AR) condition and with typical items in the typical reward (TR) condition. We introduced these exceptions to provoke memorization strategies rather than using simple rules (e.g., a single square on Dimension 1 means Humi, otherwise Tami). Overall, however, all three dimensions were similarly predictive, as indicated by nearly equal logistic regression weights.

Under the constraints described so far, we used simulations to calculate predictions using the formal models described in the preregistrations and the Modeling section. Our goal was to make sure that stimulus sets and reward manipulations led to strongly diverging behavioral predictions for the memory-strength, similarity, and attention

Table 2

Training Stimuli and Manipulations in Study 3

Food item	Dimension			Condition			Set
	1(□)	2(◇)	3(○)	C	TR	AR	
Humi							
1	1	2	3	3	30	3	TR-high
2	1	1	3	3	30	3	TR-high
3	1	4	2	3	3	30	AR-high
4[a]	4	4	1	3	3	30	AR-high
5	1	3	1	3	3	3	Neutral
Tami							
6	4	4	2	3	30	3	TR-high
7	2	2	4	3	3	30	AR-high
8	3	3	3	3	3	3	Neutral
9	2	4	2	3	30	3	TR-high
10[a]	1	2	4	3	3	30	AR-high

Note. The three dimensions were represented by the visual stimulus features (see Figure 1). The reward values (thalers) were balanced. C = control (baseline); TR = typical reward; AR = atypical reward. Final sample sizes after exclusions (see Data Cleansing) were $N = 32$ participants in the control condition, $N = 36$ participants in the TR condition, and $N = 38$ in the AR condition. Set summarizes the stimulus rewards as referred to in the Results and Modeling sections in which the sets were indexed (e.g., TR-high refers to items that rendered a high reward in the TR condition but a low reward in the AR condition).

[a] Atypical category items.

hypotheses, which we describe in more detail in Appendix A. Intuitively, this led to selecting training items with specific characteristics. That is, in Study 2, the two high-reward items (Figure 1B) had unique category-predicting dimension values (i.e., 1 and 5 on Dimension 3), which was also partially the case for Study 1 (i.e., the Humi high-reward item had unique values on Dimensions 2 and 3, and the value 5 on Dimension 1 uniquely predicted Tami for the high-reward item). Thus, it should be easy to maximize accuracy and reward for these two items. Furthermore, all high-reward items had at least one dimension with extreme values. If the similarity gradient for these items becomes broader (similarity hypothesis) this would predict categorization errors for other items due to increased similarity (and interference) of these extreme-valued high-reward exemplars.

Feature attention. We measured potential changes in weighting (or attending to) the dimensions in two ways. First, to test for changes in the absolute degree of feature importance/attention, we asked participants to estimate the relative importance of each dimension at the end of each study. They adjusted three horizontal slider bars representing the dimensions until a percentage value, which was depicted next to each slider, reflected their assessment of "relative importance" (also framed as "attention paid to each dimension") during the categorization transfer phase. The roughly 4-inch-wide slider bars moved in steps of 1 between 1 and 100 (constrained to sum to 100), and the dimensions were indexed by symbols as in Tables 1 and 2. We introduce the second way of testing cognitive changes in feature attention using a modeling approach in the Modeling section.

The stimulus characteristics described above also predict specific changes in dimension attention/importance if high-reward items receive more attention. In general, higher dimension values on average predicted Tami (see Tables 1 and 2). However, in all studies, this relation would become negative (i.e., lower dimension values predict Tami) for one dimension if attention was focused on high-reward items (e.g., Dimension 2 in Studies 1 and 2). Also, Dimension 1 in the TR condition of Study 3 would not reliably differentiate between the categories of high-reward items and thus should be considered as less important.

Reward judgments. Directly after the transfer phases of Studies 2 and 3, participants performed a reward-judgment task in the reward conditions. In Study 2, they estimated the likelihood that previously trained as well as 25 completely novel items rendered a high reward on a 9-point scale ranging from 1 (*very unlikely*) to 9 (*very likely*). The items were presented individually and in random order (see Table A4, Appendix A). The novel items were selected so that they differed in how similar they were to previously trained high-reward items. In Study 3 we individually presented all transfer items and asked participants to indicate whether they believed the item would render a high reward (yes vs. no) and how confident they were in their decision on a 7-point scale ranging from 1 (*not sure at all*) to 7 (*absolutely sure*).

In Study 3, all participants also completed an item-recognition task (deciding whether an item was seen in training or only presented in the transfer). This test was conducted after the reward judgments in the reward conditions. This made it difficult

to compare them to the baseline condition, in which no rewards were judged. Moreover, the results did not affect our conclusions and are therefore not reported (the data are provided in the Online Supplemental Material).

Participants and Procedure

For Studies 1 (online), 2 (online), and 3 (laboratory), respectively, we recruited 204, 172, and 111 participants. The studies lasted about 25, 25, and 50 min. The online studies were conducted on Amazon Mechanical Turk (Study 1: U.S. citizens, age $M = 34.97$ years, $SD = 10.32$, 93 females; Study 2: U.S. citizens, age $M = 39.88$ years, $SD = 11.53$, 90 females). The laboratory study was conducted at the University of Zurich and participants were recruited via the online database ORSEE (Greiner, 2015) and notice boards (Swiss citizens, age $M = 24.96$ years, $SD = 6.35$, 89% undergraduates, 72 females).[2]

In Study 1, participants received a lump-sum payment of \$2.00 and a bonus of \$2.66 on average (range \$1.66 to \$3.83) depending on the accumulated thalers. The exchange rates were 175 thalers and 380 thalers to \$1.00, in the baseline and the reward condition, yielding nearly equal overall payoffs in the two conditions. In Study 2, participants received a lump sum of \$2.50 instead of \$2.00, because we increased the number of training items (and trials). We also increased the contribution of the bonus, which was \$3.03 on average, ranging from \$1.30 to \$4.59, with exchange rates of 145 thalers and 350 thalers to \$1.00, in the baseline and the reward condition. In Study 3, participants received a lump sum corresponding to Swiss Franks (CHF) 16.00 per hour ($N = 69$) or course credit ($N = 42$), and a bonus payment of CHF 5.40 on average, ranging from CHF 1.85 to CHF 8.52. The exchange rates were 125 thalers and 400 thalers to CHF 1.00 in the baseline and the reward conditions.

In each study, all participants first provided informed consent and then completed a demographic survey, the training phase, and the transfer phase. In Studies 2 and 3, the reward-judgment task (reward conditions only) followed the categorization transfer phase. The feature-attention weights were surveyed at the end, before providing overall

[2]Unfortunately, there is no standard, or easy, way to calculate statistical power for (logistic) mixed-model tests (see Rights & Sterba, 2018), as this would require prior assumptions about random effect variances. Such variations are unknown, and approximate solutions have been published only for simpler designs (see Westfall, Kenny, & Judd, 2014).

performance feedback and a debriefing.

In all studies, we announced the exchange rates and the existence of a transfer phase before the training started. Participants in Studies 1 and 2 also answered two control questions, testing whether they understood the training feedback procedure and that they could increase their payoff with correct decisions. After incorrect answers, we presented a summary of the instructions and asked again, which was repeated until both answers were correct. Instead of using "catch trials" to capture inattentive participants (see Paolacci, Chandler, & Ipeirotis, 2010), we added an "honesty check" to the debriefing, where participants could self-declare whether they should not be considered for the data analysis, for example, because they had used memory tools. We made it clear that their honesty would not affect their payment.

Data Cleansing

As preregistered for the online studies, we excluded those participants who self-declared that their data should not be used (i.e., 10 and 12 participants in Studies 1 and 2, respectively), as well as those who incorrectly responded (at least) twice to the task comprehension questions (i.e., 2 and 5 participants in Studies 1 and 2, respectively). In Study 2, data from one participant were lost due to a storing error. In each study, we also excluded participants who seemed to guess at the end of the training, to ensure active learning. Instead of using classic outlier detection (e.g., Van Selst & Jolicoeur, 1994), as preregistered for Study 1, we based our inferential exclusions on Bayesian contaminant classification (e.g., Zeigenfuse & Lee, 2010; as preregistered for Study 2), justified and explained in Appendix B.

In Study 1, nine participants were classified as "guessing" in the baseline condition. In Study 2, seven participants were classified as "guessing" in each condition. In Study 3, four participants were classified as "guessing" in the baseline condition and one in the AR condition. The final sample sizes are given in Tables 1 and 2. The "guessing" exclusions were done for all subsequent analyses, which did not affect the overall (omnibus) test results but did partially affect the significance test results from the post hoc comparisons between the conditions of Studies 2 and 3. The significance tests without these contaminant exclusions can be found in the Online Supplemental Material (Categorization Accuracy).

Analysis Plan

To test our behavioral hypotheses (on accuracy) we used logistic mixed models (for an introduction, see Singmann & Kellen, in press; see also Bolker et al., 2009). Besides systematic fixed effects, mixed (or multilevel) models take the hierarchical clustering of variance in the data and effects into account (e.g., variations within groups and/or individuals). Roughly speaking, with multiple stimulus responses for each participant (fully crossed) one can separately estimate the variance between participants and between stimuli, called by-participant and by-stimulus random intercepts, respectively. Additionally, a within-participant fixed effect (e.g., of item reward) can vary in strength between participants, which is modeled as a by-participant random effect (the commonly used term random slopes is omitted to avoid confusion with the slope parameter described below). Ignoring such variance structures has been shown to lead to severe alpha-error inflation in significance testing (Judd, Westfall, & Kenny, 2012).[3]

We used Type 3 likelihood-ratio tests (in R afex::mixed; Singmann, Bolker, Westfall, & Aust, 2015), which provide p values for nonzero differences in explained variance between the full model (i.e., with all possible effects) and the restricted models (i.e., without the tested fixed effects). However, to quantify the evidence against the null hypothesis that performance is equal for each item set between conditions, we report the results of a Bayesian logistic regression (using the brms::brm package in R; Bürkner, 2017), which is hierarchically constrained on participant, stimulus, and effect variability, as done for the mixed model. We report the range of "most likely" estimates for the central parameters (i.e., 95% credible intervals, CIs; for an introduction to hierarchical Bayesian models, see M. D. Lee & Wagenmakers, 2014; Rouder & Lu, 2005), for simplicity, instead of Bayes factors (BFs; e.g., the prior-to-posterior likelihood ratio; Dickey, 1971; see also Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010, pp. 167–170, and Lee & Vanpaemel, 2018), because priors are very difficult to scale in logistic models with multiple interactions and complex variance terms (e.g., depending

[3]Please note, in contrast to the suggestion of Barr, Levy, Scheepers, and Tily (2013) to define the maximal random-effects structures that are justified by the design, we decided to define only the random structures that included the theoretically relevant variance terms (e.g., without including methodological counterbalancing variables), because overparameterization has been shown to decrease the chance of detecting existing effects (see Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017).

on variable and effect scaling and design coding; see also Rouder, Morey, Speckman, & Province, 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017).

Results

Categorization Accuracy

First, we wanted to find out whether and how categorization accuracy changes when introducing item rewards of variable magnitude, compared to an equal-reward baseline. We therefore focused on contrasting item-specific performance (i.e., trained item sets high vs. low in Studies 1 and 2, and TR-high vs. AR-high in Study 3, see Tables 1 and 2]) in the training and transfer phases. The accuracy over the training blocks (learning curves), as well as in the transfer phase, is illustrated in Figure 2, contrasting performance between the conditions (baseline vs. reward condition[s]).

As can be seen in Figure 2, varying the reward magnitude between category instances indeed seemed to affect categorization accuracy. However, instead of motivating better performance, providing a 10 times higher reward for some items seemed only to decrease accuracy on low-reward items. We tested the overall effect in each study by entering the factors of the corresponding condition (reward[s] vs. baseline), set (high/TR-high/AR-low vs. low/TR-low/AR-high), and training block (continuous, centered/standardized), including all interactions as fixed effects in the regression. In Study 3, we left out the neutral set (see Table 2) to reduce the number of tests and to focus on the manipulated (counterbalanced) items.

We assumed random intercepts for participants and stimuli, and by-participant random effects of set, training block, and their interaction. In Studies 2 and 3, we reduced the model complexity by assuming uncorrelated random intercepts and random effects, due to singularities and for reasons of model identification, respectively. We conducted a similar test in the transfer phase (trained items only), again using a logistic mixed model with condition and set and their interaction as fixed effects, with random intercepts for participants, stimuli, and transfer blocks, as well as by-participant random effects for set.

The results of the corresponding model likelihood-ratio tests in the training and transfer phases are presented in Table 3. We found significant overall interactions between the factors set and condition in Studies 1 and 3 in both training and transfer

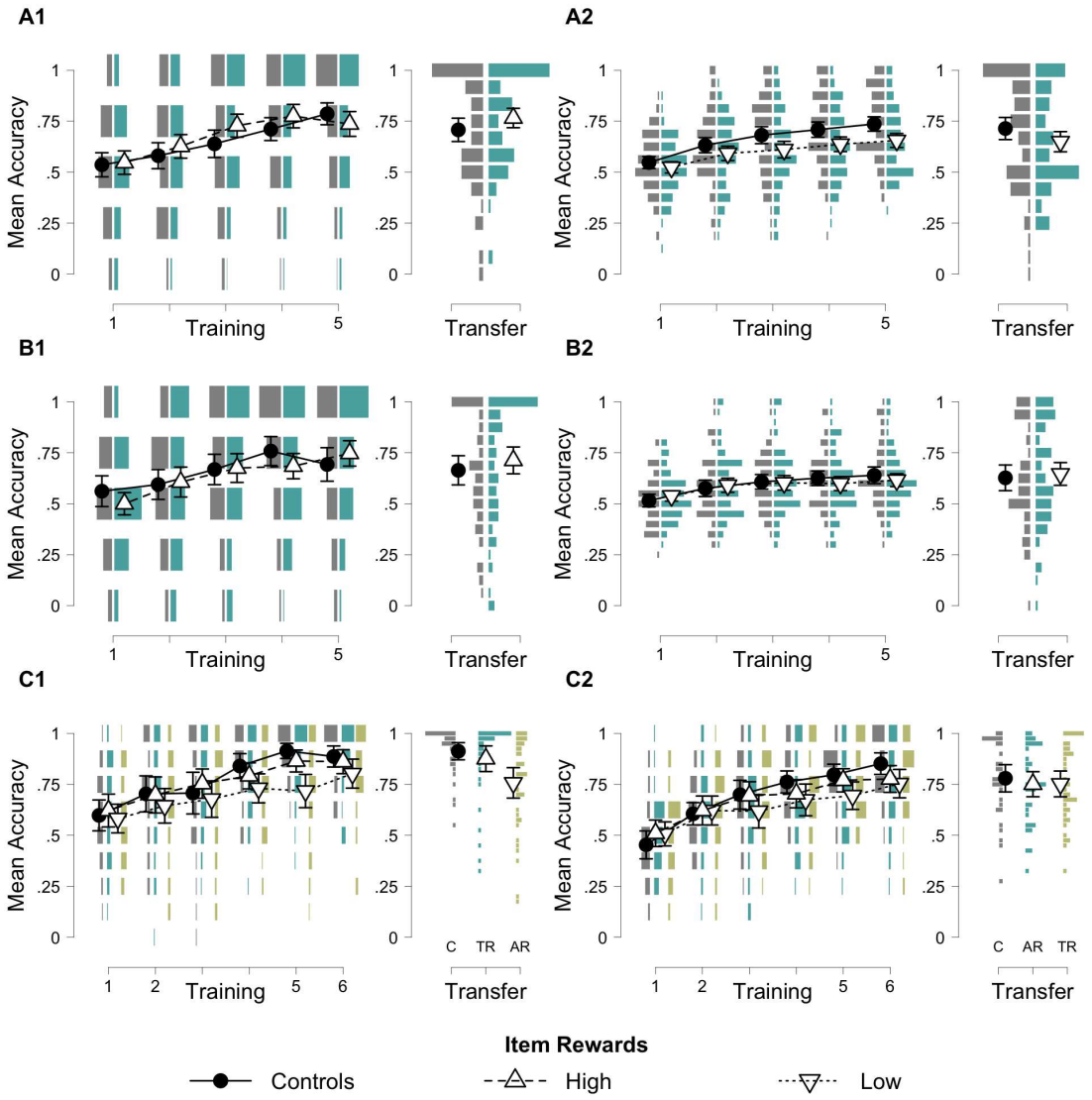


Figure 2. Categorization accuracy for (A) Study 1 (A1 = high reward; A2 = low reward), (B) Study 2 (B1 = high reward; B2 = low reward), and C() Study 3 (C1 = TR-high/AR-low; C2 = TR-low/AR-high), over blocks of training (x axes: 2 blocks cluster to 1 bin) and in the transfer phase (test). Comparison on trained items between the baseline (controls) and the reward conditions (high or low rewards; same color = same condition; see Tables 1 and 2). Error bars are 95% credible intervals of the participant means. Histograms reflect relative frequencies of individual scores, and bins reflect possible values (i.e., 0 to 1 in steps of 1/[items · trials]). C = baseline (control); TR = typical reward; AR = atypical reward.

phases, but not in Study 2. Please note, the overall learning success in Study 2 was unexpectedly low, with an average final accuracy of $M = .64$, $SE = .01$. Unfortunately,

Table 3

Categorization Accuracy. Mixed-Model Results for Training and Transfer Phases

	Effect	*df*	χ^2	*p*
Study 1	Training phase			
	Condition	1	0.26	.610
	Set	1	1.59	.207
	Block	1	121.80	**<.001**
	Condition:Set	1	7.34	**.007**
	Condition:Block	1	1.61	.204
	Set:Block	1	30.52	**<.001**
	Condition:Set:Block	1	0.74	.388
	Transfer phase			
	Condition	1	0.04	.843
	Set	1	1.25	.26
	Condition:Set	1	7.96	**.005**
Study 2	Training phase			
	Condition	1	0.10	.752
	Set	1	3.10	.073
	Block	1	105.84	**<.001**
	Condition:Set	1	0.08	.776
	Condition:Block	1	0.00	.979
	Set:Block	1	18.66	**<.001**
	Condition:Set:Block	1	2.87	.090
	Transfer phase			
	Condition	1	1.05	.306
	Set	1	2.15	.143
	Condition:Set	1	0.24	.62
Study 3	Training phase			
	Condition	2	3.41	.182
	Set	1	3.00	.083
	Block	1	131.00	**<.001**
	Condition:Set	2	8.28	**.016**
	Condition:Block	1	11.23	**.004**
	Set:Block	2	0.99	.319
	Condition:Set:Block	2	3.21	.201
	Transfer phase			
	Condition	2	7.33	**.026**
	Set	1	3.33	.068
	Condition:Set	2	6.98	**.030**

Note. Type 3 likelihood-ratio tests (full vs. restricted models). Estimates highlighted in bold significantly contribute to explained variance (.05 level).

this level of performance implies a loss of statistical power for detecting *decreases in accuracy*.

To better illustrate the source of the significant interactions, we derived the learning curve characteristics from the training data, which are usually referred to as threshold and slope in item response theory (see Baker, 2001; see also Dixon, 2008). The threshold θ defines the point on the continuous variable (i.e., training block t) at which the logistic function predicts 50% accuracy. The slope β defines the curvature of the sigmoid function over the values of the continuous variable (from slow change to S-shaped/quick change), such that the probability of being accurate in a given training block t is generally defined by $p(\text{Correct}|t) = \left(1 + \mathrm{e}^{-\beta(\theta+t)}\right)^{-1}$. The slope β, hence, is an estimate of how quickly accuracy improves over learning. The theoretically meaningful

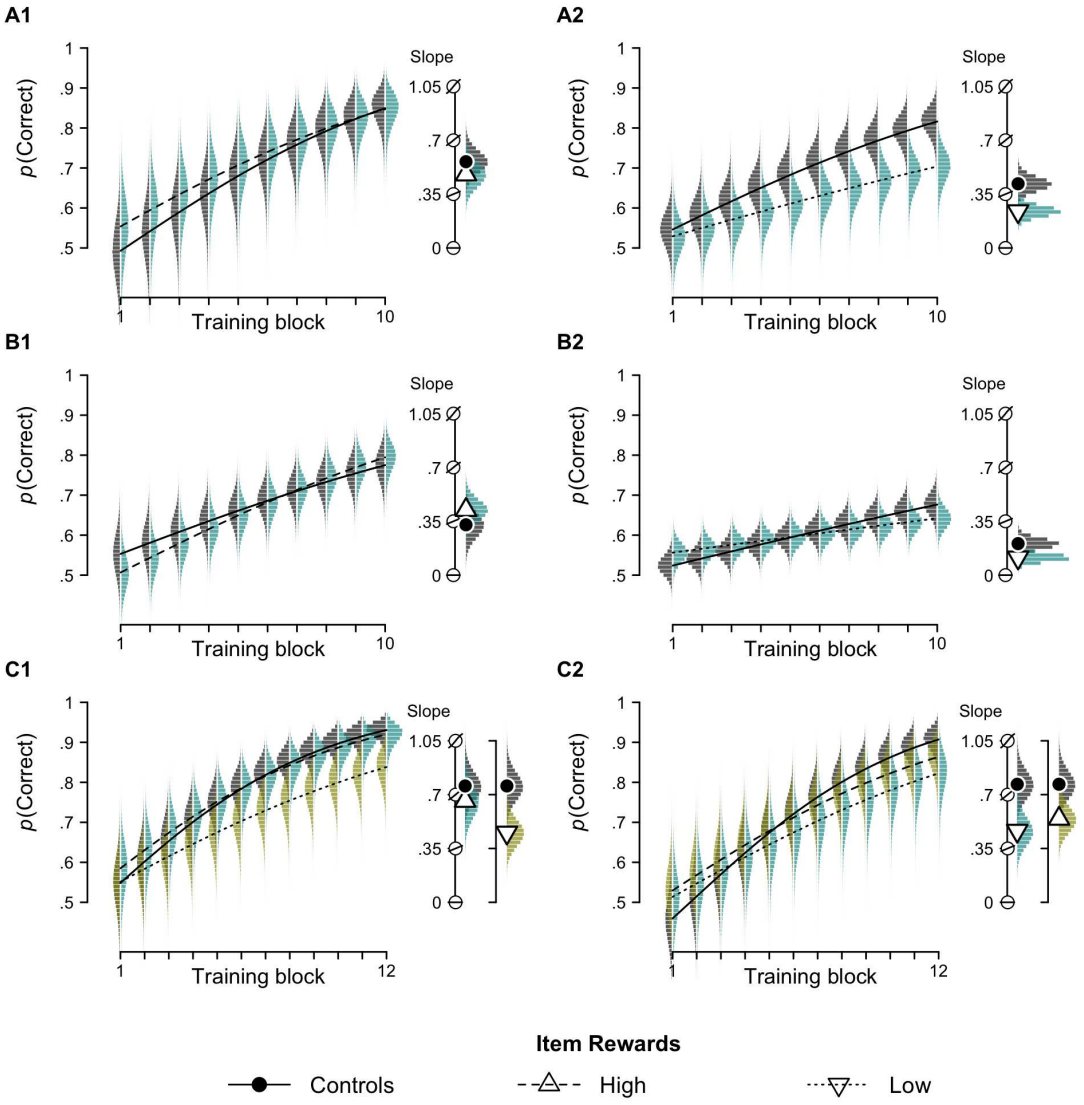


Figure 3. Predicted average accuracy (y axes) over blocks of training (x axes), based on hierarchical Bayesian regressions for (A) Study 1 (A1 = high reward; A2 = low reward), (B) Study 2 (B1 = high reward; B2 = low reward), and (C) Study 3 (C1 = TR-high/AR-low; C2 = TR-low/AR-high). The slope parameter estimates the rate of growth (log scale) toward $p(\text{Correct}) = 1$, as illustrated by the icons. Histograms depict posterior likelihoods of the means. Markers indicate average estimates. Item sets are compared between conditions (controls vs. high or low reward; same colors = same condition; see also Figure 2).

comparison between conditions is that of the slope β.[4]

[4] In the Bayesian regressions (or mixed models), the parameters are factorized as $p(\text{Correct}|t) = \left(1 + e^{-\beta\theta - \beta t}\right)^{-1}$. In regressions, $\beta\theta$ is usually referred to as the intercept, and βt is the training block

The memory-strength hypothesis and the attention hypothesis predict that accuracy improves for high-reward items while declining for low-reward items, and the similarity hypothesis predicts that accuracy declines only for low-reward items, compared to baseline. Figure 3 depicts the corresponding Bayesian estimates for the average learning curves, as well as the posterior distributions for the slope parameters (β). As we mean-centered the training block variable ($SD = 1$), the size of the slope effects can be conveniently interpreted relative to standard conventions ($0 <$ very low $< 0.35 <$ low $< 0.7 <$ moderate $< 1.4 <$ high; Baker, 2001).

We calculated the posterior differences of the slopes between the conditions on each item set from the Bayesian regression (Table 4). As can be seen, the CIs (i.e., the 95% most likely estimates) of the slope differences on low-reward items always excluded zero, even in Study 2. This was not the case for high-reward items. Please note, in Studies 1 and 2, the training parameter estimates were less certain for high-reward items than for low-reward items, due to fewer stimuli. However, the trends for high-reward items also pointed in different directions between the studies, speaking against the hypothesis of improved memory.

Finally, we also tested whether the training effects translated to the transfer phase, in which the trained items were presented together with novel items (without feedback). For the transfer phase, we implemented a logistic hierarchical Bayesian regression, defined as before, with condition and item sets as fixed effects. Again, we calculated the posterior differences in item-specific accuracy estimates (log scale) between conditions (Table 4, Transfer accuracy). As can be seen, the pattern of results changed. The learning rate effect translated to the differences in transfer phase accuracy only in Study 1 and in the AR condition of Study 3, while the CIs of all other difference estimates included zero.

In sum, the evidence generally supported the hypothesis that introducing variable stimulus rewards in category learning decreases learning accuracy for low-reward items, without leading to a benefit for high-reward items. In the transfer phase, the evidence for this specific trend seemed less strong, as it occurred in only two of the four transfer

---

multiplied by its log-linear effect. Importantly, the significant interactions between condition r and set s (i.e., systematic variations of $\beta_{rs}\theta_{rs}$), thus, shift θ to the point at which $p(\text{Correct}|t) = .5$ depending on the size of the slope (which also makes it basically impossible to define priors, e.g., for the intercept).

Table 4

Posterior Model Estimates for Effects of Reward Magnitude on Learning Rate (Training Phase) and Overall Accuracy (Transfer Phase)

Contrast	Diff	95% CI	Fig.
Training slopes			
Study 1 High	0.083	[-0.129, 0.293]	3.A1
Study 2 High	-0.096	[-0.310, 0.117]	3.B1
Study 3 High (TR)	0.097	[-0.183, 0.379]	3.C1
Study 3 High (AR)	0.224	[-0.026, 0.472]	3.C2
Study 1 Low	0.179	[0.063, 0.297]	3.A2
Study 2 Low	0.090	[0.003, 0.177]	3.B2
Study 3 Low (TR)	0.303	[0.052, 0.554]	3.C2
Study 3 Low (AR)	0.302	[0.043, 0.561]	3.C1
Transfer accuracy			
Study 1 High	-0.331	[-0.950, 0.286]	2.A1
Study 2 High	-0.450	[-1.207, 0.314]	2.B1
Study 3 High (TR)	0.101	[-0.698, 0.922]	2.C1
Study 3 High (AR)	0.106	[-0.634, 0.835]	2.C2
Study 1 Low	0.578	[0.107, 1.065]	2.A2
Study 2 Low	-0.140	[-0.677, 0.393]	2.B2
Study 3 Low (AR)	1.324	[0.577, 2.063]	2.C1
Study 3 Low (TR)	0.224	[-0.474, 0.939]	2.C2

Note. Bayesian model estimates for the differences in the learning slopes and the transfer accuracy (intercepts on log scale) between conditions on each item set (e.g., high reward vs. baseline) with 95% credible intervals (CIs). Positive values always indicate higher parameter values (e.g., steeper training gain in accuracy) in the baseline condition. Fig. references the panels in Figures 2 and 3. AR = atypical reward; TR = typical reward.

phases, which is readdressed in the Discussion and Modeling sections.

Feature Attention

Next, we examined whether the distribution of attention to the stimulus dimensions differed between the baseline and the reward conditions. Feature attention was measured via sliders, which were rescaled to p values between 0 and 1 for each dimension (summing to 1) for the analyses. The stimulus layout used in Studies 1 and 2 was identical, but in Study 2 we counterbalanced the assignment of the dimension values (with dimensions numbered as in Table 1) to the visual features to control for feature salience (see Online Supplemental Material, Feature Attention for descriptive details).

To test for systematic changes in the dimension ratings (coded as in Tables 1 and 2) we conducted Dirichlet regressions (R, DirchletReg; Maier, 2014), estimating the mean and precision parameters of the three-dimensional probability distribution, with condition as a predictor in each study. A significant effect of condition would reflect an

overall change in the three dimension weights induced by our reward manipulations (i.e., the model with separate Dirichlet distributions by condition explains significantly more variance than a model with only one distribution). In Study 3, we also entered an indicator for the six feature-counterbalancing conditions (main effect and interaction with condition; see Online Supplemental Material for further details).

In short, there were no significant effects of our manipulations on the distribution of reported dimension weights/feature attention. However, we replicated an attention bias of about 45% on average given to the most salient berry dimension in both Studies 1 and 2 (also found when inferring dimension attention via cognitive modeling, see Modeling section), suggesting that the measure captured meaningful variance. In fact, the reported attention values differed between the reward and baseline conditions by less than 5% in all comparisons. Thus, variable rewards in category learning did not seem to systematically affect feature importance in the categorization transfer task. To cross-validate these results we also analyzed the attention estimates derived from the model analyses in the Modeling section. However, these results led to the same conclusions and therefore are detailed only in the Feature Attention section in the Online Supplemental Material.

Discussion

In general, reward magnitude did, indeed, affect item-specific categorization accuracy. We found evidence in all studies and conditions that low-reward items were learned more slowly compared to the same items in an equal-reward baseline condition. In contrast, in none of the studies did we find evidence that providing 10 times higher reward for some of the training items was beneficial for learning these items, compared to the baseline, even if it would have been easy to exploit their unique characteristics (Study 2) to maximize accuracy and payoff. There was no evidence that introducing variable stimulus rewards affected the attention paid to each dimension.

We further found that the effects of reward magnitude on learning did not always persist into the corresponding transfer phases, which suggests that encountering novel items may reduce the effect of reward and possibly changes the categorization process. It is especially noteworthy that a carryover occurred only in one of the two reward conditions in Study 3. Although the learning rate was equally affected in both reward

conditions, it carried over (very strongly) to the transfer phase of the AR condition, but not of the TR condition. As the labels imply, in the AR condition high reward was paired with atypical category items (or exceptions), while in the TR condition high reward was paired with typical category items.

In Study 2, one major shortcoming was the unexpectedly low overall performance, with about 60–65% final training accuracy. This might be due to the dimension values more strongly overlapping between the categories, compared to Studies 1 and 3, leading to more confusion. Nonetheless, although performance was generally low, we still found evidence for slower learning (slope) of low-reward items, which was relatively weak compared to Studies 1 and 3, but confidently larger than zero. As pointed out, even within this rather difficult task, participants did not or could not exploit the fact that the high-reward items had unique feature values, which could have considerably increased their payoff.

In sum, the behavioral analyses provide some striking evidence that variable stimulus rewards might be detrimental when learning about stimulus categories. One caveat when interpreting this effect is, however, that the hypotheses depend on the assumption that all participants rely on the same categorization processes (e.g., exemplar-based inference or cue abstraction). As this assumption is often doubted (e.g., Ashby et al., 2002; Bröder et al., 2010; Donkin et al., 2015; Hoffmann et al., 2016), we obtained further evidence regarding the relation between reward magnitude and the different potential categorization processes, by carrying out more fine-grained analyses using computational modeling.

Modeling

We initially introduced the hypotheses that reward magnitude could affect the availability of exemplars in memory (memory-strength hypothesis), the similarity gradients of high-reward exemplars (similarity hypothesis), or the distribution of attention to stimulus dimensions (attention hypothesis). The behavioral trends, so far, challenge the memory-strength and attention hypotheses, as they predicted an increase in accuracy for high-reward items compared to baseline. The predicted between-participant differences follow from expected behavioral changes within participants. The observed within-participant changes (i.e., the difference in

performance for high- and low-reward items), however, is a qualitative prediction that all three hypotheses would make.

One possibility to differentiate the hypotheses on a within-person level is to use cognitive modeling. Cognitive models formalize the assumptions of psychological theories, such as similarity-based or rule-based category inference. Based on estimates of free (unknown) cognitive parameters, such as feature attention or exemplar similarity gradients, each cognitive model makes point predictions for behavioral data, which makes it possible to test them against each other via likelihood measures. The between-condition effect we found (a detriment for low-reward items, without a benefit for high-reward items) was predicted by the similarity hypothesis, which was derived from the idea that the similarity gradients of high-reward exemplars might become broader. Hence, when analyzing the transfer phase using cognitive models that represent the three hypotheses one would expect that a GCM (Nosofsky, 1986, 2011) with exemplar-specific similarity gradients is more likely than a GCM version with exemplar-specific memory strengths, representing the memory-strength hypothesis (please note that attention varies freely in both models). Furthermore, there should be systematic differences in exemplar similarity gradients on high-reward items between reward conditions and the baseline condition. Additionally, the mixed behavioral transfer results might be explained by a mixture of strategies. If the similarity hypothesis best describes the behavior of one population of participants, the remaining population might not show changes in cognitive processing. For instance, for the cue-abstraction model the estimated dimension weights (feature attention) should be equal between the conditions, to be consistent with the reported behavioral data on accuracy.

In the following subsection we explain how we disentangle potential effects on exemplar memory strength and similarity gradients in the GCM (Medin & Schaffer, 1978; Nosofsky, 1986), which has been successfully used to describe how people categorize and how they judge item familiarity and typicality (Hahn & Chater, 1997; Jäkel et al., 2008a; Jäkel, Schölkopf, & Wichmann, 2008b; Lamberts, 1995; G. Murphy, 2004; Nosofsky, 1988a, 1988b, 1992; Nosofsky, Clark, & Shin, 1989; Nosofsky & Palmeri, 1997; Pothos, 2007; Pothos & Bailey, 2000; Pothos & Wills, 2011; Sigala, Gabbiani, & Logothetis, 2002; von Helversen, Herzog, & Rieskamp, 2014).

The Generalized Context Model

The GCM (using Nosofsky's, 2011, parameter notation) is formally based on the assumption that decision makers retrieve instance representations from memory (stored exemplars j, e.g., truffle and morel vs. yellow knight and death cap) to evaluate their similarity s_{ij} to the current stimulus i. In its basic form (Nosofsky, 1986) the model derives the category probability (e.g., edible or poisonous) for a given stimulus $p(Q|i)$ as the overall similarity to exemplars of that category relative to the similarities to all available exemplars in q categories.

$$p(Q|i) = \frac{\sum_{j \in Q} s_{ij}}{\sum_q (\sum_{j \in q} s_{ij})} \tag{1}$$

The similarities s_{ij} are obtained, first, by calculating the difference between the stimulus i and each exemplar j as the sum of psychological distances between the feature values x_{im} and x_{jm} across the feature dimensions m (e.g., color and shape).

$$d_{ij} = \sum_m (w_m |x_{im} - x_{jm}|^r)^{1/r} \tag{2}$$

The parameter r defines the metric for the dimension distances (e.g., city block with $r = 1$), which are then weighted by the attention w_m to the feature dimensions. An attention weight shrinks or expands the distances between a stimulus and the exemplars on each dimension. On the one hand, a higher value of w_m increases the sensitivity to small differences on that dimension m, and categorizations should become more accurate for (e.g., high-reward) exemplars that discriminate on this dimension. On the other hand, if an attention weight increases on one dimension it automatically decreases for other dimensions, because the weights are defined to sum to 1. Differences on dimensions with zero weights are neglected, which makes categorizations less accurate for (e.g., low-reward) exemplars that are dissimilar on those dimensions.

After summing the weighted dimension distances, the subjective similarity s_{ij} is then obtained via exponentiation:

$$s_{ij} = \mathrm{e}^{-c d_{ij} + \ln(V_j)} \tag{3}$$

The psychological distance is weighted by the similarity gradient c and added to $\ln(V_j)$, which describes an exemplar's memory strength.[5] First, with narrow similarity

[5] Please note, this formula is equivalent to the original (i.e., $V_j \mathrm{e}^{-c d_{ij}} = \mathrm{e}^{-c d_{ij} + \ln(V_j)}$) but more clearly reveals the formal nature of memory strength as a similarity intercept/bias on the log scale.

gradients (i.e., large values of c), exemplars become less similar to distant stimuli more quickly than with broad similarity gradients (i.e., low values of c). In the decision process, thus, narrow gradients represent (precise) integration of those exemplars close to the stimulus (and strong stimulus-category discrimination), and broad gradients represent integration of more distant exemplars (and more probabilistic predictions). The self-similarity of an exemplar is always 1, meaning that c describes an indirect influence of exemplars on stimuli. The c parameter, therefore, is often interpreted as an individual's sensitivity to (dis)similarity, but sometimes also in terms of confusion, because broader gradients predict more categorization (identification) errors than precise gradients (for details, see Jäkel et al., 2008b; Miendlarzewska et al., 2016; Nosofsky, 2011).

Second, however, the multiplicative influence of the weighted exemplar distance can be overruled by the memory strength of an exemplar, which can be viewed as a retrieval rate. That is, large values of V_j not only increase the contribution of an exemplar to Equation 1 but also indirectly reduce the contribution of other exemplars, independent of their similarity to the stimulus. For instance, as initially described, if two high-reward exemplars have higher memory strengths than all other exemplars, this "erases" the low-reward exemplars from Equation 1, turning the GCM into a reward-prototype model (following a "first recalled, first served" principle).

Finally, the similarity gradient can be allowed to vary between high- and low-reward exemplars (c_j). If the gradient becomes broader (i.e., c decreases) for high-reward exemplars in the reward conditions, this would lead to a stronger integration across a wider range of stimulus space. This increases their weight in the decision process without becoming more available. This means low-reward exemplars become more similar to the high-reward exemplars relative to themselves, which introduces interference and, hence, decision errors. This interference might have an even stronger effect on accuracy if the high-reward exemplars are atypical for their category, as in Study 3 (AR).

Crucially, the within-participant tendency to categorize high-reward exemplars more accurately than low-reward exemplars can be predicted by changes in exemplar memory strength but also by changes in their similarity gradients, or in dimension-specific attention. However, the similarity hypothesis predicts that this stems

from a decrease in accuracy for low-reward exemplars, instead of an increase in accuracy for high-reward exemplars. This means each mechanism potentially explains the individual data, but their descriptive adequacy further depends on more complex predictions regarding category generalization for novel items, which we pin down using Bayesian cognitive modeling.

Hierarchical Bayesian Modeling of Categorizations

To test which mechanism describes the influence of reward magnitude in exemplar-based categorization decisions, we applied two separate versions of the GCM (i.e., a memory and a similarity version; Nosofsky, 2011) to all decisions from the transfer phase in each study and condition, describing the influence of reward magnitude either by variations on item-specific memory or by the similarity gradient, respectively.[6] Figure 4 shows the graphical model of the similarity version of the GCM, with prior definitions for Study 3. We tested the two GCMs against each other in a hierarchical Bayesian latent mixture framework (see M. D. Lee & Wagenmakers, 2014; p. 212, for an introduction).

To distinguish the exemplar account (the GCM) from cue abstraction (CAM) we included a standard (mean-centered/unbiased) logistic regression (see also Juslin, Jones, et al., 2003). Additionally, we implemented a recognition and bias model designed to capture a family of unsystematic response strategies. Specifically, this model freely estimates accuracy for trained items (varying between recognition and guessing), as well as a choice bias for the novel items. A full description of all models and priors can be found in Appendix C; here, we focus on the crucial aspects.

One of the four latent models was assigned to each participant, using a hierarchical transdimensional Markov chain Monte Carlo (MCMC) method (see Sisson, 2005). In this method, a categorical indicator selected the most likely model for a participant (k) in each iteration. This indicator z_k was hierarchically drawn from a Dirichlet distribution (ζ_q^z), reflecting the mixture model likelihoods in each condition q (see Bartlema, Lee, Wetzels, & Vanpaemel, 2014; M. D. Lee & Vanpaemel, 2008; for similar

[6] Please note, a GCM that includes free parameters for memory and similarity for the same exemplars is very difficult to identify since they can cancel each other out (see Equation 3). Or as Jäkel et al. (2008a) put it in a related discussion: "The exemplar weights defeat the purpose of introducing a similarity measure for the stimuli" (p. 266).

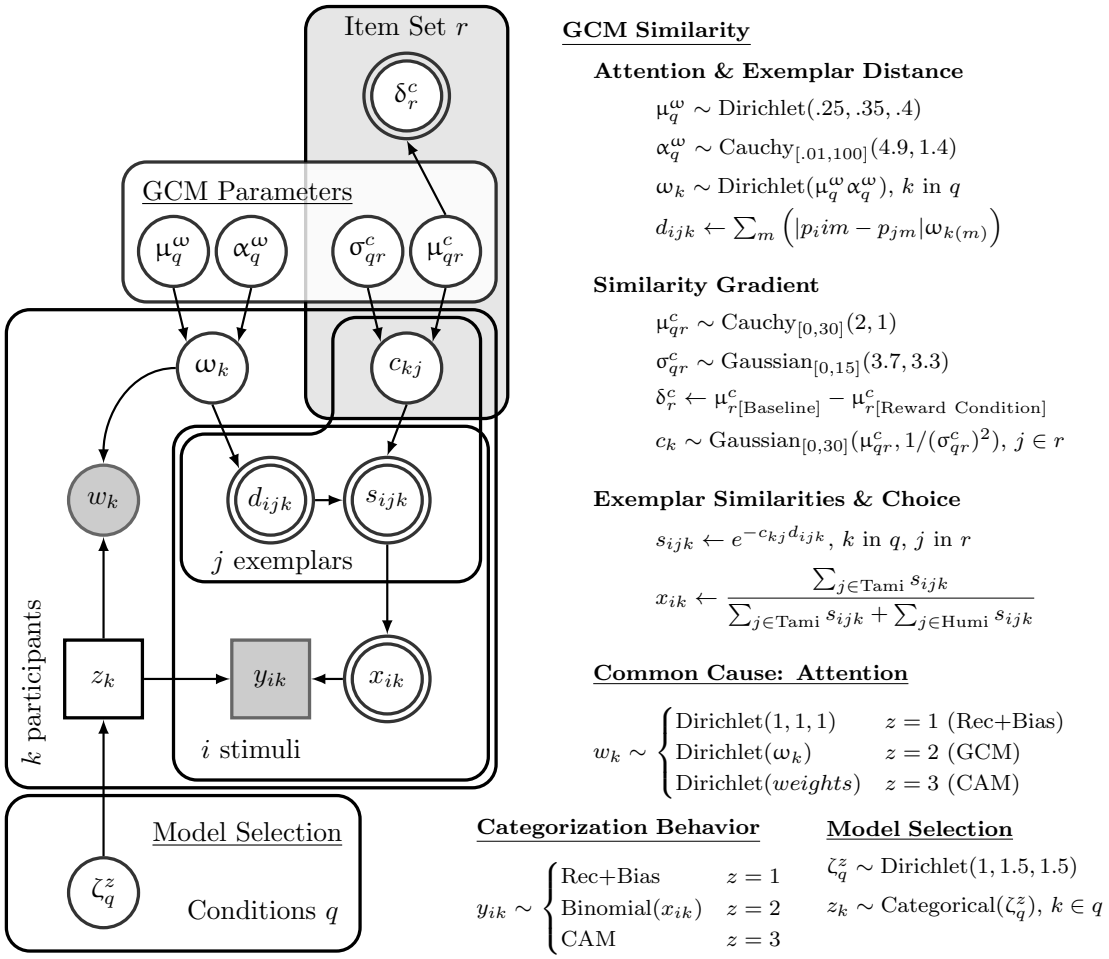


Figure 4. Graphical Bayesian model of the generalized context model (GCM) with separate similarity gradients μ_{qr}^c for each set r of exemplars (indexed according to set in Tables 1 and 2). Prior definitions are given for Study 3. The shaded box highlights the tested difference δ_r^c of the estimated similarity gradients μ_{qr}^c on each set r between the reward condition and the corresponding baseline. Circles and squares represent continuous and nominal variables, respectively. The model is embedded in a model selection, and a common cause mechanism predicts both reported feature weights w_k and the total number of Tami responses y_{ik} (gray symbols) for each item i in the transfer phase; see text).

applications and related discussions). The frequency (across sampling) with which a model is selected for a participant was used as a measure for the confidence in the individual assignment. To address potential issues raised by some researchers (e.g., Sisson, 2005)—namely, that unlikely priors can affect the model and parameter

likelihoods and thereby bias model selection—we implemented pseudo (data-driven) priors. These priors were based on fits of the single models only, each with uninformed priors (except for reasonable parameter constraints). Using different priors in the model selection, however, led to very similar results.

Simultaneously, we tested the parameter differences between the conditions for the memory strengths (V_j) and the similarity gradients (c_j) of the j exemplars. For this, we hierarchically estimated the corresponding model hyperparameters (i.e., population means and standard deviations) for the manipulated item sets r (indexed according to Tables 1 and 2) in each condition (q). Specifically, in the memory GCM, we estimated μ_{qr}^{V} for the item sets that differed on reward between conditions (i.e., set r = [high] in Studies 1 and 2, and sets r = [TR-high/AR-low] and r = [TR-low/AR-high] in Study 3). We fixed the memory parameter to $\ln(V_j) = 0$ (see Equation 3) for items that yielded the same (low) reward in all conditions (i.e., set r = [low] in Studies 1 and 2, and set r = [neutral] in Study 3).[7]

In the similarity GCM (Figure 4; without fixed values), we hierarchically sampled the exemplar gradients c_{kj} from the estimated population distribution with mean μ_{qr}^{c} and standard deviation σ_{qr}^{c} (see Appendix C for details). As outlined, we would expect the differences between the similarity estimates in the q = baseline condition to indicate narrower generalization than those in the q = reward condition(s) for r = high-reward exemplars (reflected by $\delta_r^c > 0$ in Figure 4), but to remain unchanged for r = low-reward exemplars.

However, because exemplar-specific parameters substantially increase the flexibility of the GCM in all cases, we also included a common-cause constraint, to improve model convergence and theoretical plausibility. In particular, each of the cognitive models implements feature attention/weighting parameters (e.g., ω_k in Figure 4), which are essential for making predictions. Thus, the models' performance evaluations were based on both the categorizations and the subjective feature-attention measures (w_k in Figure 4). For the CAM, in which the dimension weights were not constrained to being positive and sum to 1 (as in the GCM), we normalized the estimated *absolute* weights for fitting

[7] By definition of the GCM, a model that fixes memory strength for high-reward items and estimates memory strength for low-reward items is formally identical to the reported model. A model with free memory strengths for high- and low-reward items is unidentifiable (see Appendix C).

the reported weights. Please note, negative weights in the CAM (ceteris paribus) reverse the polarity of the corresponding dimension but leave the common interpretation unchanged (i.e., indicating the sensitivity to each dimension in predicting choice). For the recognition and bias model, we assumed random equal weights, for convenience.

Modeling Results and Discussion

The model ran on six chains with 100,000 iterations in all studies and with sampler adaptation periods of 20,000 iterations (R package JAGS; Plummer, 2015). However, there were severe issues with the memory GCM: Specifically, the chains for μ_{qr}^{V} did not converge. All attempts to solve this issue failed (e.g., reparameterization, more samples). Furthermore, across all studies, only 10 participants were confidently assigned to this model (z_k frequency>90%), without showing evidence for the hypothesized effect on memory strength. A GCM memory model alone (single model fit) fully converged but also did not show evidence for the hypothesized memory effect. These analyses suggest that the memory GCM captured rather singular (and unreliable) behavioral patterns during the model selection procedure, but not the observed effects of our reward manipulations, which is consistent with the conclusions from the results on categorization accuracy.

Because of these convergence issues and to be able to reliably estimate the evidence for the similarity hypothesis, we conducted a second analysis excluding the memory GCM. In this analysis, all models fully converged (same settings as before). The individual model assignments in Table 5 show that the GCM was the dominant model in Study 3, but on par with the CAM in Studies 1 and 2 (see Appendix Table D1 for all parameter estimates, and Appendix Figure D1 for average categorization behavior in each group). To evaluate the mapping between the predictions and the data, we derived the posterior predictions (see M. D. Lee & Wagenmakers, 2014) of the GCM and CAM. That is, for both model populations we simulated category predictions for each transfer item, taking all uncertainty of the estimated hyperparameters into account. These posterior predictive checks (see Figures 2 and 3 in the Online Appendix on OSF) clearly indicate systematically different predictions between the models, with an imperfect but overall satisfactory match between the actual behavior of the participants and the model predictions.

Table 5

Number of Participants Best Fit by Each Model

Condition	Rec/Bias	GCM	CAM
Study 1			
Baseline	11 (9)	43 (39)	36 (32)
Reward	17 (15)	41 (38)	32 (30)
Study 2			
Baseline	13 (12)	22 (17)	26 (20)
Reward	15 (12)	31 (27)	34 (20)
Study 3			
Baseline	13 (12)	16 (15)	3 (3)
Typical reward	10 (10)	21 (19)	5 (4)
Atypical reward	16 (12)	18 (15)	4 (3)

Note. Numbers in parentheses reflect participants assigned with at least 90% confidence (i.e., a model was more likely than all other models in more than 90% of the Markov chain Monte Carlo samples). Rec/Bias = Recognition and bias model; GCM = generalized context model; CAM = cue-abstraction model.

Additionally, about 35% of the participants were classified as being in the recognition and bias group in Study 3, but only 16% and 20% of the participants in Studies 1 and 2, respectively. Please note that an assignment to this model group can mean that a participant's behavior is either poorly described by the GCM or CAM, or well described by assuming guessing or a choice bias (or both). In fact, this group showed systematic behavior (e.g., good recognition of trained instances) in Study 3, but not in Study 1 or 2 (see Figure 1 in the Online Appendix on OSF). However, given that the cognitive processes underlying these responses are unclear, we focus here on the GCM and CAM groups.

To test the hypothesis that a higher reward alters exemplar similarity gradients (i.e., $\delta_r^c \neq 0$) we first derived the distribution for the null hypothesis by calculating the differences between two half-Cauchy priors as defined for μ_{qr}^c (see also Figure 4). To approximate the likelihood at zero difference we took the posterior density in the [-0.1, 0.1] interval for the null distribution, as well as for the corresponding parameter posteriors for each item set. Please note, due to the MCMC model selection procedure, the participant group assignments determine which participants contribute (and how much) to the estimation of these posteriors (i.e., the GCM posteriors only update for participants k for which $z_k = 2$, see Figure 4).

Table 6

Reward-Induced Differences in Exemplar-Specific Similarity

	Similarity difference		
Contrast	δ_r^c	95% CI	BF
Low reward (Set)			
Study 1	1.01	[-0.37, 2.42]	1.35
Study 2	0.11	[-2.07, 2.36]	0.62
Study 3: TR (neutral)	-0.03	[-3.63, 3.47]	0.96
Study 3: AR (neutral)	-1.61	[-5.38, 1.87]	1.33
Study 3: AR	1.12	[-1.79, 4.04]	1.13
Study 3: TR	-1.29	[-4.13, 1.59]	1.14
High reward (Set)			
Study 1	2.08	[0.34, 3.87]	9.58
Study 2	3.41	[0.06, 8.11]	6.98
Study 3: TR	0.63	[-2.09, 3.4]	0.87
Study 3: AR	4.96	[2.15, 7.93]	>100

Note. Posterior means for exemplar-similarity differences (δ_r^c in Figure 4) between the baseline and each reward condition (see sets in Tables 1 and 2) in the generalized context model group. Positive values reflect broader similarity gradients in the reward condition compared to the baseline. In Study 3, The neutral set rendered 3 thalers in all conditions, and the high-reward set in the atypical reward (AR) condition (which was the low-reward set in the typical reward [TR] condition) included two atypical items. CI = Credible interval; BF = Bayes factor.

We then calculated the BFs (SD density ratios; Dickey, 1971), which are presented together with the 95% CIs in Table 6. As can be seen, except for the TR condition in Study 3, the similarity gradients for high-reward exemplars were confidently smaller than the corresponding baseline estimates (CIs > 0, and BFs > 6.9). But there was no evidence of differences in the measures for low-reward items (all BFs ~ 1) in all studies and conditions.

In sum, the model analyses, indeed, supported the hypothesis that the similarity gradients of high-reward instances become broader, compared to baseline, if participants rely on exemplar-based category inference. This seemed to be the case even in Study 2, in which we found no effect of reward magnitude on accuracy in the transfer phase. In contrast, in the TR condition of Study 3, in which we also did not find evidence for an effect of reward magnitude on accuracy in the transfer phase, the GCM analysis did not suggest broader similarity gradients for high-reward items.

While the mixed pattern of results in Study 2 seems related to a combination of a

generally weak performance and a latent mixture of cognitive processes (GCM and CAM) in the transfer phase, the pattern in Study 3 still remains unclear and might also depend on whether participants actually learned (or generalized) the reward values of the stimuli. Thus, we conducted a final analysis of the participants' reward judgments, to test how well they recognized the reward values of the trained category instances, and to investigate the relationship between reward generalization to novel items and category generalization.

Predicting Reward Judgments From Categorizations

In Studies 2 and 3, participants completed reward-judgment tasks in the reward conditions after the category-transfer phase. They had to predict whether correctly classifying a presented item would result in a high or a low reward. If similar (or interacting) cognitive processes underlie exemplar-based category inference and reward-value inference, we would assume that both can be described by exemplar similarity gradients. Accordingly, the GCM parameters that best describe the categorization decisions should also allow predicting the participants' reward judgments, at least as long as they have learned which training items have high and low rewards.

To test this assumption, we first derived predictions for the reward judgments based on exemplar similarity using the GCM; that is, we calculated the probability (which we then treated as a judgment) that a presented item will be classified as high reward by providing the model with the exemplars' outcome labels (high vs. low reward) instead of their response labels (Tami vs. Humi). Importantly, to make a priori predictions for the reward judgments, we used the GCM parameter posteriors that were estimated in the category-transfer phase (see Appendix Table D1), instead of fitting the judgment data.

More precisely, in each iteration of the model selection for the category-transfer decisions, we simultaneously presented each item of the reward-judgment task to the GCM. The prediction of the probability of a high reward was based on a parameter sample from the hyperposteriors. The averages of the resulting distributions of item-judgment predictions, thus, reflect the expected average item-reward judgments of the GCM population. We compared the predicted average judgments to the actual item judgments, aggregated within each group of most confidently classified participants (i.e.,

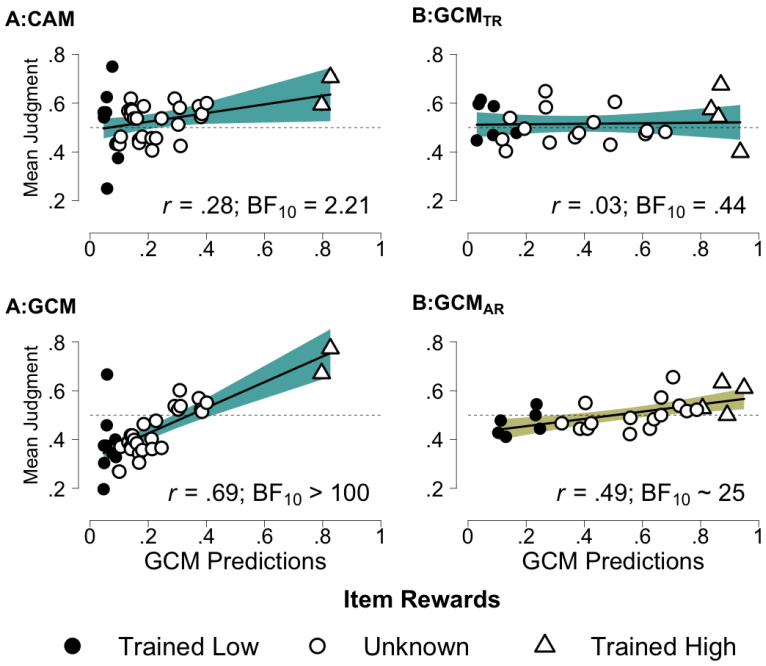


Figure 5. Reward-value generalization. The item-specific predictions of the GCM (x axes) are plotted against the corresponding reward judgments (y axes) for Studies 2 (A) and 3 (B). For Study 2, A:CAM and A:GCM refer to the subgroups of participants inferred in the reward condition (see Table 5). For Study 3, B:GCM$_{TR}$ and B:GCM$_{AR}$ refer to the participants in the two reward conditions (see Table 2). Color coding corresponds to Figures 2 and 3. Markers depict single items. Unknown refers to untrained transfer (Study 3) or novel (Study 2) items. Shaded areas are standard regression confidence intervals (95%). Bayes factors (BFs) reflect evidence for $r \neq 0$ (see text). TR = Typical reward; AR = atypical reward.

with confidence > .9 in Table 5). To avoid making assumptions about the mapping between the judgment and the prediction scales, we focused on correlational evidence instead of other fit measures.[8]

The relation between the item judgments and the GCM predictions are depicted in Figure 5. The correlations were obtained using Bayesian estimation (R package

[8]Although conceptually equivalent, this approach deviates from the preregistered analysis, for which we planned to extend the Bayesian framework (transfer phase) to jointly estimate the GCM parameters for the categorizations and the reward judgments. The presented approach was chosen for simplicity, and to ease the interpretation of the GCM judgment predictions of the different groups.

BayesFactor::correlationBF; Morey & Rouder, 2018; medium prior scale = 1/3, undirected). As can be seen, in Study 2 there was very strong evidence of a correlation between the GCM predictions and the reward judgments of the GCM group (Figure 5A:GCM, CI_r [0.50, 0.83]), while there was only weak evidence of correlations when participants were assigned to CAM (Figure 5A:CAM, CI_r [-0.01, 0.54]). The general pattern within the recognition and bias group (not shown) was nearly identical to that in the CAM group, with $r = .3$, BF = 2.7, CI_r [0.0, 0.56]).

Thus, in Study 2, similarity-based reward generalization was evident for those participants who also seemed to rely on instance similarity in category inference. Please note, most low-reward items (black circles in Figure 5A:GCM) were judged as being likely low-reward items in this group, suggesting that both high- and low-reward items were accurately recognized. Crucially, about 48% of the variance in the average reward judgments was predicted by the GCM with the parameter values estimated in the categorization task. In contrast, the scatter plots for the CAM group in Study 2 (Figure 5A:CAM) show that while participants in this group still tended to correctly respond to high-reward items (triangles), the judgments for novel items (white circles) and trained low-reward items (black circles) are rather randomly distributed around 50%.

Finally, in Study 3 there was no evidence of a linear relation between the average item judgments and the GCM predictions in condition TR (Figure 5B:GCM_{TR}, CI_r [-0.32, 0.38]), but strong evidence in condition AR (Figure 5B:GCM_{AR}, CI_r [0.17, 0.73]). This pattern mirrors the observed effects for categorization accuracy, and the model analyses in the transfer phases, suggesting the possibility that the absent effect of reward magnitude in the transfer phase of the TR condition is related to the absence of reward-value generalization in the GCM group. Together with Study 2, the overall pattern leaves little doubt that reward generalization influences category generalization, if instance similarity drives both types of inferences, whereas categorization processes based on cue-abstraction processes do not seem to be affected.

General Discussion

Reward magnitude has been argued to be the driving force guiding preferential decision making (Kahnt, 2018; Schultz, 2006; Sutton & Barto, 1998). Additionally, stimuli associated with higher reward values automatically attract more attention than

those associated with lower reward values in various cognitive tasks, such as working memory and perceptual decision tasks (e.g., Allen & Ueno, 2018; B. A. Anderson, 2013; Della Libera & Chelazzi, 2009; Klink et al., 2017; Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009; Miendlarzewska et al., 2016). These perspectives suggest that variable rewards for correct decisions might also influence cognitive processes in inferential decision tasks and, thus, objective decision accuracy. We investigated this question in a category-learning paradigm. We compared three hypotheses for how variable rewards might affect categorizations by affecting (a) memory for high-reward exemplars, (b) their psychological stimulus similarity, or (c) the distribution of attention to stimulus dimensions (see Le Pelley et al., 2016; Miendlarzewska et al., 2016; Seger & Peterson, 2013).

In three studies we tested the related behavioral hypotheses, whether introducing variations in stimulus-specific reward magnitude for correct decisions increases categorization accuracy for high-reward items and decreases accuracy for low-reward items. For this, we compared item-specific scores for categorization accuracy in the category-learning and transfer phases of conditions with variable stimulus rewards (high vs. low) to conditions in which all stimuli were associated with the same reward.

Overall, against the common intuition that providing a higher reward motivates (or enhances) learning, none of the three studies provided evidence that high-reward items were learned systematically better than in the baseline condition with equal rewards, even if high-reward items had unique identifying features. In contrast, we found that learning speed and accuracy were reliably impeded for low-reward items (i.e., during learning) in all studies and conditions. Furthermore, in two of the three studies, this reduction in categorization accuracy persisted in a subsequent category-transfer phase.

Although general performance differences between high- and low-reward items (i.e., within participants) are compatible with the memory-strength, similarity, and attention hypotheses, only the similarity hypothesis predicts that these differences reflect a detriment for low-reward items (i.e., compared to the baseline condition). In contrast, the memory-strength and attention hypotheses predicted a concurrent increase in accuracy for high-reward items. The similarity hypothesis received further support in a computational modeling approach using hierarchical Bayesian model selection. We tested the three cognitive assumptions, as implemented in an exemplar (GCM;

Nosofsky, 2011) and a cue-abstraction account (CAM, using the weighted sum of feature values; e.g., Juslin, Jones, et al., 2003) on their ability to predict the categorizations of the participants in the transfer phase.

Specifically, for those participants best described by the GCM, we found moderate to strong evidence that reward magnitude changed the similarity gradients of high-reward exemplars (becoming broader), compared to baseline. We found this effect in all three studies, except for one condition (discussed below). However, the results speak against the idea that a higher reward enhances exemplar memory strength or that reward magnitude leads to a shift in attention to dimensions that predict category membership of high-reward items.

To shed further light on the relationship between reward learning and category generalization, we also asked the participants to complete a reward-judgment task. In it, they estimated the likelihood that presented items would render a high versus low reward. We found that these judgments correlated with the predictions of a GCM model adapted to predict reward magnitude. Most crucially, the reward-judgment predictions were obtained using the most likely GCM estimates from the category-transfer phase. However, we found a correlation only for the group of participants who were assigned to an exemplar-based strategy in the category-transfer phase, and not for CAM participants.

In sum, the evidence from our studies supports the idea that reward magnitude impedes category learning for low-reward instances without a benefit for high-reward instances, most probably by affecting processes of similarity-based inference. In the following, we discuss the theoretical implications of these results, as well as boundary conditions, limitations, and future directions.

Reward, Memory, and Similarity

At first glance, our finding that high-reward items are not categorized more accurately, compared to baseline, seems to contradict previous research reporting that reward enhances stimulus recall in working memory tasks (or memory strength). It is important to note, however, that the idea of reward-boosted memory (e.g., Wallis et al., 2015) usually refers to a difference between high- and low-reward stimuli. Thus, it seems worth considering a detriment for low-reward stimuli as a potential explanation, as well.

The general pattern of our category-learning results, thus, is in line with previous findings on the effect of reward magnitude on cognitive processes such as memory and attention (e.g., Allen & Ueno, 2018; B. A. Anderson, 2013; Della Libera & Chelazzi, 2009; Klink, Jentgens, & Lorteije, 2014; Klink et al., 2017). But instead of the memory-strength hypothesis, they support the hypothesis that prioritizing high-reward stimuli affects their similarity gradients during exemplar-based decisions (i.e., broader gradients of high-reward instances). However, the key aspect of this interpretation—an absence of reward-enhanced learning—was partially based on inconclusive evidence for the null hypothesis. Thus, this result requires further investigations that include a broader set of category structures.

Nonetheless, a recent study similarly showed that presenting all stimuli more often—usually assumed to increase memory strength (see also Nosofsky, 2011)—does not increase the accuracy with which stimuli are classified, compared to a control condition (Homa et al., 2018). The authors proposed that the influence of stimulus frequency on categorization, as well, might be better described by varying similarity gradients, rather than memory strength.

Indeed, memory strength seems to be a meaningful formal concept to capture frequency effects in tasks that explicitly ask for recalling or recognizing stimuli (e.g., Donkin et al., 2015; Nosofsky, 1991b). However, stimulus frequency also seems to affect categorizations differently from what might be expected from a GCM perspective (e.g., Hendrickson et al., 2019; Homa et al., 2018). In line with Nosofsky's (1991) argument that empirical investigations need to substantiate theorized principles, these findings raise the general question of whether categorization phenomena that seem intuitively related to memory strength (e.g., reward and frequency) might rather point to changes in stimulus-specific similarity gradients (see also Goldstone & Son, 2005; Hendrickson et al., 2019). More critically, it has been argued that exemplar memory weights are prone to model overfitting while annulling the meaning of similarity gradients in exemplar-learning models (Jäkel et al., 2008a; p. 266).

Thus, assuming changes in similarity gradients might be a more general path to explain effects of reward magnitude or stimulus frequency. In fact, by changing their "attraction" as a function of similarity (instance-specific similarity gradients), exemplars can become "magnets" in stimulus space, a metaphor that has been used to

describe the behavioral effects of stimulus frequency in categorization (Nosofsky, 1991b; p. 114). However, the similarity perspective still has a caveat. Although, the idea that reward magnitude affects similarity gradients seems to apply in both categorization decisions and reward-learning tasks (Kahnt et al., 2012), the way it does so is inconsistent. While the behavioral results of Kahnt et al. (2012) indicate narrower (or more precise) gradients for high-reward items, we estimated broader (less precise) gradients for high-reward items in the categorization task with two exclusive categories.

As mentioned in the Introduction, this apparent contradiction also seems to be present under manipulations of stimulus frequency, as studied by Hendrickson et al. (2019). Their data suggest that the effect of increased stimulus frequency also depends on whether two exclusive categories are learned, or only one category (A vs. not-A). That is, the categorization behavior in these conditions seems compatible with broader and narrower similarity gradients, respectively.

Last, we found that the reward judgments in Studies 2 and 3 could be predicted by the GCM based on the most likely model parameters in the prior categorization task. This suggests that similarity-based category generalization and similarity-based reward generalization are closely related and are potentially equally affected by reward magnitude. This dovetails with the idea that similarity-based generalization in categorization is a process tightly linked to or moderated by the similarity-based generalization of stimulus reward value (Miendlarzewska et al., 2016; Seger & Peterson, 2013).

Insights from neuroscience corroborate this idea, showing that one common mechanism affects both similarity processes, defined by the pattern of dopaminergic activity, which is induced by corrective feedback and rewards. Reward magnitude seems to modulate this activity (see Berridge, Robinson, & Aldridge, 2009; Everitt, Morris, O'Brien, & Robbins, 1991; Galvan et al., 2005; Knutson, Adams, Fong, & Hommer, 2001; Mahler & Berridge, 2009; Miendlarzewska et al., 2016), also in category learning (see Daniel & Pollmann, 2010). Increased dopaminergic activity, in turn, might affect how stimulus features are associated with the categories, and the notion of inference based on associative memory often goes hand in hand with the idea of similarity-based inference (e.g., Kahnt et al., 2012; Miendlarzewska et al., 2016; Wimmer et al., 2012; Wimmer & Shohamy, 2012; see also Seger & Peterson, 2013, p. 1190).

Reward and Dimension Attention

Neither participants' self-reports nor the computational modeling indicated that participants changed which stimulus dimensions they attended to in response to stimulus-specific rewards. This is somewhat surprising in light of research suggesting that people direct their attention to stimulus features that best predict the desired outcomes (e.g., Le Pelley et al., 2016; Sutton & Barto, 1998). One simple reason why this assumption did not generalize to our category-learning task could be that obtaining (high) rewards required participants to categorize the stimuli accurately. Nevertheless, our attention hypothesis built on the general idea that attention shifts to feature dimensions that differentiate between the categories of high-reward items (predicting enhanced performance for these items). Of course, such hypotheses depend on the theoretical framework used to define the circumstances of observing a positive effect of reward magnitude.

Thus, the null effect of variable rewards on dimension attention deserves further empirical investigation (e.g., using different categorization tasks and reward structures), including theoretical questions about the fundamental processes that drive attention learning. One possibility that would allow integrating our results with common reinforcement-learning theories would be that dimensional attention mainly depends on which criterion/response is learned in a task, rather than its more or less desired consequences (i.e., people learn to predict either category membership or reward, or other labels). This would imply that stimulus-specific reward magnitude also might not influence dimension attention when a reward is delivered regardless of categorization accuracy, which seems an interesting avenue for further research.

Boundary Conditions, Rules, and Exceptions

Although our studies present a coherent pattern of results, they also indicate that variable rewards did not always lead to systematic changes in cognitive processing, compared to the baseline condition. First, when participants were best described by a CAM, there was no evidence that their dimension weighting was affected by reward magnitude. It seems consistent that in Study 2, these participants also did not generalize the reward magnitude of the training stimuli to novel ones and also failed to recognize low-reward stimuli.

Thus, in line with the above discussion, it seems that reward generalization and its influence on category learning is limited to similarity-based decision processes, and it should be noted that we used task structures that arguably increase reliance on similarity-based processing, that is, so-called information integration category structures, in all three studies (e.g., Ashby et al., 2002; Bröder et al., 2010; Donkin et al., 2015; Hoffmann et al., 2016; Juslin, Olsson, & Olsson, 2003). It seems worth investigating whether and how such category and task characteristics (including stimulus design) might influence the pattern of results, for instance, by more-or-less "inducing" participants to use rule-based strategies. In this vein, notwithstanding the descriptive capabilities of the GCM (Nosofsky, 1986), it seems inconsistent how the model describes reward effects via similarity gradients, because it requires the assumption that the direction of the effect depends on the task type (as in "one vs. two categorization tasks"; see Hendrickson et al., 2019; Kahnt et al., 2012; see also Polk et al., 2002, Schechtmann, Laufer, & Paz, 2010). From a more general perspective, this raises the question of whether there is a more consistent way to describe these findings (e.g., via learning models), which might also affect their cognitive interpretations (e.g., similarity vs. memory strength) which we address in future research.

Second, in Study 2 the detrimental effect of reward magnitude was seemingly present during category learning but was obscured by overall low performance, providing a natural constraint on observing this effect. Moreover, if decision strategies moderate the influence of reward magnitude on categorizations, reward effects may depend on a mixture of cognitive processes in the population. In line with this idea, we did not observe an overall effect of reward magnitude on accuracy in the category-transfer phase in Study 2, but the model analyses revealed evidence that reward affected the similarity gradients of high-reward exemplars for the subgroup of participants best described by the GCM.

Last, differences between the *typicality* of high- and low-reward items could have influenced whether the effects of reward magnitude occurred. While we found the strongest effect of reward magnitude on similarity gradients in the AR condition of Study 3, in which we paired high reward with atypical category exemplars, the effect was absent in the counterbalancing condition TR, in which we paired high reward with typical exemplars. One potential explanation for the strong effect in the AR condition

could be that integrating atypical exemplars should lead to more errors because they are not representative of their categories. Thus, increasing the similarity gradient of atypical items might enhance this effect even further.

The absence of an effect in the TR condition, however, seems more puzzling and could be related to the presence of atypical items in general or less effective reward learning, as it appeared that reward generalization was concurrently absent in the GCM group of the TR condition. However, given the exploratory nature of Study 3, and the ongoing debate regarding the special role of atypical items (or exceptions) in memory and their implications for the representation of categories (e.g., Erickson & Kruschke, 1998; Nosofsky et al., 1994; Poldrack et al., 2001; Poldrack & Foerde, 2008; Savic & Sloutsky, 2019; Schlegelmilch, Wills, & von Helversen, 2018), these explanations should be considered speculative, and further research is needed to disentangle the effects of typicality and item rewards.

Limitations, Implications, and Future Directions

Finally, it is also important to consider the limitations and future directions of our research approach. First, although the behavioral effects during learning indicate that reward affects learning processes, our modeling analyses focused on the transfer phase decisions. Thus, our analyses provide the insight that similarity-based processes may be involved in the interaction between reward and category learning but do not rule out alternative cognitive approaches. Consequently, our findings highlight the need to investigate existing category-learning accounts, such as the configural cue model (e.g., Gluck & Bower, 1988), the attention learning covering map (ALCOVE) (Kruschke, 1992), the supervised and unsupervised stratified adaptive incremental network (SUSTAIN) (Love et al., 2004), the rational model (Sanborn et al., 2010), the divergent autoencoder (Kurtz, 2007), the S.O.S. network (Goldstone et al., 1996), the category abstraction learning model (Schlegelmilch et al., 2018), and other accounts (see Pothos & Wills, 2011) and their assumptions about how reward magnitude affects learning and generalization, and to test their ability to consistently explain empirical data in different tasks.

The second limitation concerns our focus on positive reward values. The observed effects might not directly translate to negative values (punishment; but see Schechtman,

Laufer, & Paz, 2010). For instance, Kahnt (2018, p. 328) discussed that values (or magnitudes) are neurally coded on a common scale (see also Kahnt, Park, Haynes, & Tobler, 2014), but that the valence of reinforcement (positive vs. negative) is distinctly processed. Furthermore, dopaminergic activity might serve different functions, that is, serving to motivate or inhibit depending on whether there is a reward or punishment (e.g., Bromberg-Martin, Matsumoto, & Hikosaka, 2010; Schultz, 2006). In addition, an important further question is whether the influence of reward magnitude on cognitive processing is moderated by individual differences in reward (or reinforcement) sensitivity (e.g., Corr, 2004; Daniel & Pollmann, 2010).

Despite these limitations, the possibility that differences in reward magnitude, which are omnipresent in everyday life, might impede the cognitive processes of category learning bears importance from two general perspectives. For one, category learning shares several aspects with other cognitive domains, such as working memory (Jonides et al., 2008; Lewandowsky, 2011; Oberauer et al., 2007), inferential choice, and probabilistic decision making (e.g., Achtziger et al., 2015; Pachur & Olsson, 2012), and any task that includes the memorization and identification of stimuli and their multiattribute integration (for further discussions, see Goldstein, 1991; Juslin, Jones, et al., 2003; Markman & Ross, 2003; Miendlarzewska et al., 2016; Russell et al., 1999; Seger et al., 2015; Seger & Peterson, 2013; E. R. Smith & Zarate, 1992). A common understanding of the influence of reward magnitude might help identify the converging theoretical constructs and unify our understanding of their underlying cognitive processes.

Second, it needs to be shown if reward magnitude similarly affects learning and generalization of stimulus categories in more natural categorization or judgment tasks (see G. L. Murphy, 2016; J. D. Smith, 2005). This includes considerations of how natural categories are structured, and how reward magnitude is distributed within and between categories. For instance, in the stock market, rare instances might tend to be more profitable, and also more similar to each other, while being riskier (e.g., belonging to different behavioral response categories).

On a broader scale, a potential influence of reward magnitude on stimulus similarity seems to concern any domain in which similarity is considered as a viable theoretical model of information integration (see also Polk et al., 2002). This includes

the domain of economic decision making but also therapeutic training, machine learning, and neuroscience, which, as well, increasingly focus on the described mechanisms of stimulus generalization and reinforcement to understand learning and behavior (see Jäkel et al., 2008a, 2008b; Miendlarzewska et al., 2016; O'Doherty et al., 2017; Swan, Carper, & Kendall, 2016; for related reviews, respectively).

Conclusion

Taken together, our results suggest that introducing instance-specific differences in reward magnitude impedes learning of category instances associated with low rewards, compared to equal-reward control conditions, without being beneficial to high-reward instances. We conducted hierarchical Bayesian model analyses on category-transfer decisions after learning, using an exemplar model (GCM; Nosofsky, 1986, 2011; Medin & Schaffer, 1978) and a CAM (weighted additive; e.g., N. Anderson, 1981; Juslin, Jones, et al., 2003), which included making predictions for an additional reward-value generalization task. The combined results suggest that when participants rely on exemplar memory or similarity-based processes during category inference, but not when they rely on cue abstraction, learning and generalizing stimulus reward values can be reliably predicted as well. Crucially, when relying on exemplar similarity to infer the categories of novel stimuli, the similarity gradients of high-reward stimuli seem to become broader, which also predicts increased confusion errors among low-reward instances. Cognitive functions of cue abstraction (i.e., cue validity or feature attention) seem unaffected by stimulus-specific rewards.

Acknowledgments

This project was funded by a grant from the Swiss National Science Foundation to the second author (No. 157432). The first author wants to thank the supervisors and participants in the Summer School for Mathematical and Computational Modeling (2016), especially Chris Donkin and Klaus Oberauer, and in the workshop Bayesian Modeling for Cognitive Science (2017), especially Michael D. Lee and Eric J. Wagenmakers. The authors further want to thank Andy Wills, Klaus Oberauer, and three anonymous reviewers for highly valuable comments on earlier versions of the manuscript. We would like to thank Anita Todd for copy editing the manuscript.

Context of Research

Our research is situated within a broader research program on exemplar memory processes in judgment and decision making and was motivated by the open question of how performance rewards can influence category representations and decision behavior. We initially assumed that increasing the reward value for correctly classifying specific stimuli would increase the quality with which they are memorized, but none of our studies supported this idea. After the first study (Study 3), we focused on testing (falsifying) this reward-enhancement effect (Studies 1 and 2) under optimal conditions for observing it. We simultaneously made sure we would be able to test an important theoretical alternative that might be related to the replicated finding, that introducing variable rewards in category learning rather impedes learning of low-reward items. Consequently, the discussed theoretical implications will guide our future research focusing on conceptual replications in various categorization tasks and on finding novel category-learning accounts that can consistently integrate related phenomena.

References

Aberg, K. C., Müller, J., & Schwartz, S. (2017). Trial-by-trial modulation of associative memory formation by reward prediction error and reward anticipation as revealed by a biologically plausible computational model. *Frontiers in Human Neuroscience*, *11*, 1–15. https://doi.org/10.3389/fnhum.2017.00056

Achtziger, A., Alós-Ferrer, C., Hügelschäfer, S., & Steinhauser, M. (2015). Higher incentives can impair performance: Neural evidence on reinforcement and rationality. *Social Cognitive and Affective Neuroscience*, *10*(11), 1477–1483. https://doi.org/10.1093/scan/nsv036

Allen, R. J., & Ueno, T. (2018). Multiple high-reward items can be prioritized in working memory but with greater vulnerability to interference. *Attention, Perception, & Psychophysics*, *80*(7), 1731–1743. https://doi.org/10.3758/s13414-018-1543-6

Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of Vision*, *13*(3), 1–16. https://doi.org/10.1167/13.3.7

Anderson, N. (1981). *Foundations of information integration theory.* New York, NY: Academic Press.

Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M. A., & Poldrack, R. A. (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology*, *92*(2), 1144–1152. https://doi.org/10.1152/jn.01209.2003

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. https://doi.org/10.1037/0278-7393.14.1.33

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*(5), 666–677. https://doi.org/10.3758/BF03196423

Baker, F. B. (2001). *The basics of item response theory.* ERIC Document Reproduction Service No. ED99CO0032. Retrieved from `https://files.eric.ed.gov/fulltext/ED458219.pdf`

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for

confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Skrull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition, vol. 3content and process specificity in the effects of prior experiences* (Vol. 3, pp. 61–88). Hillsdale, NJ: Erlbaum.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150. https://doi.org/10.1016/j.jmp.2013.12.002

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015, Jun). Parsimonious Mixed Models. *arXiv e-prints*, arXiv:1506.04967.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward:'Liking','wanting', and learning. *Current Opinion in Pharmacology*, *9*(1), 65–73. https://doi.org/10.1016/j.coph.2008.12.014

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, *112*(2), 330–336. https://doi.org/10.1016/j.cognition.2009.04.008

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Brehmer, B. (1976). Social judgment theory and the analysis of interpersonal conflict. *Psychological Bulletin*, *83*(6), 985–1003. https://doi.org/10.1037/0033-2909.83.6.985

Bröder, A., Gräf, M., & Kieslich, P. J. (2017). Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision Making*, *12*(5), 491–506.

Bröder, A., Newell, B. R., & Platzer, C. (2010). Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: A matter of cue representation. *Judgment and Decision Making*, *5*(5), 326–338.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in

motivational control: Rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834. https://doi.org/10.1016/j.neuron.2010.11.022

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chen, L., Meier, K. M., Blair, M. R., Watson, M. R., & Wood, M. J. (2013). Temporal characteristics of overt attentional behavior during category learning. *Attention, Perception, & Psychophysics*, *75*(2), 244–256. https://doi.org/10.3758/s13414-012-0395-8

Corr, P. J. (2004). Reinforcement sensitivity theory and personality. *Neuroscience & Biobehavioral Reviews*, *28*(3), 317–332. https://doi.org/10.1016/j.neubiorev.2004.01.005

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *The Quarterly Journal of Experimental Psychology*, *65*(3), 439–464. https://doi.org/10.1080/17470218.2011.608854

Daniel, R., & Pollmann, S. (2010). Comparing the neural basis of monetary reward and cognitive feedback during information-integration category learning. *Journal of Neuroscience*, *30*(1), 47–55. https://doi.org/10.1523/JNEUROSCI.2205-09.2010

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. https://doi.org/10.1037/h0037613

De Leeuw, J. R. (2015). jspsych: A Javascript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Della Libera, C., & Chelazzi, L. (2006). Visual selective attention and the effects of monetary rewards. *Psychological Science*, *17*(3), 222–227. https://doi.org/10.1111/j.1467-9280.2006.01689.x

Della Libera, C., & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychological Science*, *20*(6), 778–784. https://doi.org/10.1111/j.1467-9280.2009.02360.x

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223. https://doi.org/10.1214/aoms/1177693507

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933–948. https://doi.org/10.1037/xlm0000083

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. https://doi.org/10.1037/0096-3445.127.2.107

Everitt, B., Morris, K., O'Brien, A., & Robbins, T. (1991). The basolateral amygdala-ventral striatal system and conditioned place preference: Further evidence of limbic-striatal interactions underlying reward-related processes. *Neuroscience*, *42*(1), 1–18. https://doi.org/10.1016/0306-4522(91)90145-E

Galvan, A., Hare, T. A., Davidson, M., Spicer, J., Glover, G., & Casey, B. (2005). The role of ventral frontostriatal circuitry in reward-based learning in humans. *Journal of Neuroscience*, *25*(38), 8650–8656. https://doi.org/10.1523/JNEUROSCI.2431-05.2005

Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A. (2009). Introduction: A brief history of neuroeconomics. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 1–12). London, England: Academic Press. https://doi.org/10.1016/B978-0-12-374176-9.00001-4

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247. https://doi.org/10.1037/0096-3445.117.3.227

Goldstein, W. M. (1991). Decomposable threshold models. *Journal of Mathematical Psychology*, *35*(1), 64–79. https://doi.org/10.1016/0022-2496(91)90034-Q

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 13–36). New York, NY: Cambridge University Press.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual*

Conference of the Cognitive Science society (pp. 243–248).

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 43–92). Hove, England: Psychology Press.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344–354. https://doi.org/10.1037/0278-7393.7.5.344

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102. https://doi.org/10.1016/j.cogpsych.2019.03.001

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*(6), 2242–2261. https://doi.org/10.1037/a0037989

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1193–1217. https://doi.org/10.1037/xlm0000241

Homa, D., Blair, M., McClure, S. M., Medema, J., & Stone, G. (2018). Learning concepts when instances never repeat. *Memory & Cognition*, *47*(3), 395–411. https://doi.org/10.3758/s13421-018-0874-9

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008a). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271. https://doi.org/10.3758/PBR.15.2.256

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008b). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297–303. https://doi.org/10.1016/j.jmp.2008.03.001

Jenkins, G. D., Mitra, A., Gupta, N., & Shaw, J. D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology*, *83*(5), 777–787. https://doi.org/10.1037/0021-9010.83.5.777

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*(4), 482–553. https://doi.org/10.1016/S0010-0285(02)00505-4

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193–224. https://doi.org/10.1146/annurev.psych.59.103006.093615

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 924–941. https://doi.org/10.1037/0278-7393.29.5.924

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

Kahnt, T. (2018). A decade of decoding reward-related fMRI signals and where we go from here. *Neuroimage*, *180*, 324–333. https://doi.org/10.1016/j.neuroimage.2017.03.067

Kahnt, T., Park, S. Q., Burke, C. J., & Tobler, P. N. (2012). How glitter relates to gold: Similarity-dependent reward prediction errors in the human striatum. *Journal of Neuroscience*, *32*(46), 16521–16529. https://doi.org/10.1523/JNEUROSCI.2383-12.2012

Kahnt, T., Park, S. Q., Haynes, J.-D., & Tobler, P. N. (2014). Disentangling neural representations of value and salience in the human brain. *Proceedings of the National Academy of Sciences*, *111*(13), 5000–5005. https://doi.org/10.1073/pnas.1320189111

Klink, P. C., Jentgens, P., & Lorteije, J. A. (2014). Priority maps explain the roles of

value, attention, and salience in goal-oriented behavior. *Journal of Neuroscience*, *34*(42), 13867–13869. https://doi.org/10.1523/JNEUROSCI.3249-14.2014

Klink, P. C., Jeurissen, D., Theeuwes, J., Denys, D., & Roelfsema, P. R. (2017). Working memory accuracy for multiple targets is driven by reward expectation and stimulus contrast with different time-courses. *Scientific Reports*, *7*(1), 9082. https://doi.org/10.1038/s41598-017-08608-4

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*(16), RC159. https://doi.org/10.1523/JNEUROSCI.21-16-j0002.2001

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295x.99.1.22

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576. https://doi.org/10.3758/BF03196806

Lafond, D., Lacouture, Y., & Cohen, A. L. (2009). Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychological Review*, *116*(4), 833–855. https://doi.org/10.1037/a0017188

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, *124*(2), 161–180. https://doi.org/10.1037/0096-3445.124.2.161

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron*, *64*(3), 431–439. https://doi.org/10.1016/j.neuron.2009.09.040

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities, and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*, *32*(8), 1403–1424. https://doi.org/10.1080/03640210802073697

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive

models. *Psychonomic Bulletin & Review*, *25*(1), 114–127. https://doi.org/10.3758/s13423-017-1238-3

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*(10), 1111–1140. https://doi.org/10.1037/bul0000064

Levine, L. J., & Edelstein, R. S. (2009). Emotion and memory narrowing: A review and goal-relevance approach. *Cognition and Emotion*, *23*(5), 833–875. https://doi.org/10.1080/02699930902738863

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720–738. https://doi.org/10.1037/a0022639

Lewandowsky, S., Yang, L.-X., Newell, B. R., & Kalish, M. L. (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 881-904. https://doi.org/10.1037/a0027298

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Mackintosh, N. J. (1974). *The psychology of animal learning.* New York, NY: Academic Press.

Maddox, W. T., & Bohil, C. J. (2001). Feedback effects on cost-benefit learning in perceptual categorization. *Memory & Cognition*, *29*(4), 598–615. https://doi.org/10.3758/BF03200461

Mahler, S. V., & Berridge, K. C. (2009). Which cue to "want?" central amygdala opioid activation enhances and focuses incentive salience on a prepotent reward cue. *Journal of Neuroscience*, *29*(20), 6500–6513. https://doi.org/10.1523/JNEUROSCI.3875-08.2009

Maier, M. J. (2014). DirichletReg: Dirichlet regression for compositional data in R. Retrieved from `https://epub.wu.ac.at/4077/`

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613. https://doi.org/10.1037/0033-2909.129.4.592

Matsuka, T., & Corter, J. E. (2008). Observed attention allocation processes in category learning. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1067–1097. https://doi.org/10.1080/17470210701438194

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 333–352. https://doi.org/10.1037/0278-7393.10.3.333

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(4), 241–253. https://doi.org/10.1037/0278-7393.7.4.241

Miendlarzewska, E. A., Bavelier, D., & Schwartz, S. (2016). Influence of reward motivation on human declarative memory. *Neuroscience & Biobehavioral Reviews*, *61*, 156–176. https://doi.org/10.1016/j.neubiorev.2015.11.015

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-4.2)

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, *23*(4), 1035–1042. https://doi.org/10.3758/s13423-015-0834-3

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65. https://doi.org/10.1037/0278-7393.14.1.54

Nosofsky, R. M. (1991a). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*(1), 94–140. https://doi.org/10.1016/0010-0285(91)90004-8

Nosofsky, R. M. (1991b). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53. https://doi.org/10.1146/annurev.ps.43.020192.000325

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization.* Cambridge University Press.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 282–304. https://doi.org/10.1037/0278-7393.15.2.282

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315. https://doi.org/10.1037/a0022494

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

https://doi.org/10.1037/0033-295X.101.1.53

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation*, *51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49–75). New York: Oxford University Press.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, *68*, 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. https://doi.org/10.1016/j.cogpsych.2012.03.003

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Plummer, M. (2015). Jags version 4.0. 0 user manual. Retrieved from `https://sourceforgenet/projects/mcmc-jags/files/Manuals/4x`

Poldrack, R. A., Clark, J., Paré-Blagoev, E. A., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550. https://doi.org/10.1038/35107080

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience & Biobehavioral Reviews*, *32*(2), 197–205. https://doi.org/10.1016/j.neubiorev.2007.07.007

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: Asymmetries induced by manipulating exposure frequency. *Cognition*, *82*(3), B75–B88. https://doi.org/10.1016/S0010-0277(01)00151-2

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, *133*(2), 227–244. https://doi.org/10.1037/0033-2909.133.2.227

Pothos, E. M., & Bailey, T. M. (2000). The role of similarity in artificial grammar

learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 847–862. https://doi.org/10.1037/0278-7393.26.4.847

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511921322

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407. https://doi.org/10.1016/0010-0285(72)90014-X

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. https://doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811–829. https://doi.org/10.1037/0278-7393.31.5.811

Rights, J. D., & Sterba, S. K. (2018). Quantifying explained variance in multilevel models: An integrative framework for defining r-squared measures. *Psychological Methods*, *24*(3), 309–338.

Rodrigues, P. M., & Murre, J. M. (2007). Rules-plus-exception tasks: A problem for exemplar models? *Psychonomic Bulletin & Review*, *14*(4), 640–646. https://doi.org/10.3758/BF03196814

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304–321. https://doi.org/10.1037/met0000057

Russell, G. J., Ratneshwar, S., Shocker, A. D., Bell, D., Bodapati, A., Degeratu, A., ...

Shankar, V. H. (1999). Multiple-category decision-making: Review and synthesis. *Marketing Letters*, *10*(3), 319–332. https://doi.org/10.1023/A:1008143526174

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Savic, O., & Sloutsky, V. M. (2019). Assimilation of exceptions? Examining representations of regular and exceptional category members across development. *Journal of Experimental Psychology: General*, *148*(6), 1071–1090. https://doi.org/10.1037/xge0000611

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience*, *30*(31), 10460–10464. https://doi.org/10.1523/JNEUROSCI.2377-10.2010

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2018, July). CALM–A process model of category generalization, abstraction and structuring. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 2436–2441). Austin, TX: Cognitive Science Society.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, *57*, 87–115. https://doi.org/10.1146/annurev.psych.56.091103.070229

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*(1), 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Seger, C. A., Braunlich, K., Wehe, H. S., & Liu, Z. (2015). Generalization in category learning: The roles of representational and decisional uncertainty. *Journal of Neuroscience*, *35*(23), 8802–8812.

Seger, C. A., & Peterson, E. J. (2013). Categorization= decision making+ generalization. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1187–1200. https://doi.org/10.1016/j.neubiorev.2013.03.015

Seger, C. A., & Spiering, B. J. (2011). A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, *5*, 66.

https://doi.org/10.3389/fnsys.2011.00066

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464. https://doi.org/10.3758/PBR.17.4.443

Shohamy, D., Myers, C., Kalanithi, J., & Gluck, M. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 219–236. https://doi.org/10.1016/j.neubiorev.2007.07.008

Sigala, N., Gabbiani, F., & Logothetis, N. (2002). Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience*, *14*(2), 187–198. https://doi.org/10.1162/089892902317236830

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2015). afex: Analysis of factorial experiments. *R package version 0.13–145*. Retrieved from `https://CRAN.R-project.org/package=afex`

Singmann, H., & Kellen, D. (in press). An introduction to linear mixed modeling in experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology.* Psychology Press. Retrieved from `http://singmann.org/download/publications/singmann_kellen-introduction-mixed-models.pdf,preprint`

Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, *100*(471), 1077–1089. https://doi.org/10.1198/016214505000000664

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3–21. https://doi.org/10.1037/0033-295X.99.1.3

Smith, J. D. (2005). Wanted: A new psychology of exemplars. *Canadian Journal of Experimental Psychology*, *59*(1), 47–53. https://doi.org/10.1037/h0087460

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436. https://doi.org/10.1037/0278-7393.24.6.1411

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.*

Cambridge, MA: MIT Press.

Swan, A. J., Carper, M. M., & Kendall, P. C. (2016). In pursuit of generalization: An updated review. *Behavior Therapy*, *47*(5), 733–746. https://doi.org/10.1016/j.beth.2015.11.006

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York, NY: Macmillan.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642–1645. https://doi.org/10.1126/science.1105370

Tricomi, E., & Fiez, J. A. (2008). Feedback signals in the caudate reflect goal achievement on a declarative memory task. *Neuroimage*, *41*(3), 1154–1167. https://doi.org/10.1016/j.neuroimage.2008.02.066

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*(4), 732–749. https://doi.org/10.3758/PBR.15.4.732

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. https://doi.org/10.1080/14640749408401131

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental Psychology*, *61*, 12–22. https://doi.org/10.1027/1618-3169/a000221

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wallis, G., Stokes, M. G., Arnold, C., & Nobre, A. C. (2015). Reward boosts working memory encoding over a brief temporal window. *Visual Cognition*, *23*(1-2), 291–312. https://doi.org/10.1080/13506285.2015.1013168

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, *35*(7), 1092–1104. https://doi.org/10.1111/j.1460-9568.2012.08017.x

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273. https://doi.org/10.1126/science.1223252

Wolosin, S. M., Zeithamova, D., & Preston, A. R. (2012). Reward modulation of hippocampal subfield activation during successful associative encoding and retrieval. *Journal of Cognitive Neuroscience*, *24*(7), 1532–1547. https://doi.org/10.1162/jocn_a_00237

Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362. https://doi.org/10.1016/j.jmp.2010.04.001

Appendix A

Stimulus Selection and Model Simulations

To find the optimal design, first, we randomly sampled training items according to the predefined general category structures and constraints described in the main article (Tasks, Manipulations, and Stimuli) and randomly assigned a high or low reward to some of them (balanced between categories; e.g., one high-reward item in each category of Study 1). In Studies 1 and 3, we used the sample items to map our hypotheses onto parameter variations in the GCM (i.e., 4 times higher memory strength for high-reward exemplars [representing a stronger integration of high-reward exemplars] vs. equal memory strengths for all exemplars [representing the baseline]) and a CAM (i.e., a regression with dimension weights based on the training stimuli but including four additional copies of the high-reward stimuli [representing a stronger focus on high-reward items] vs. equal weight regression [representing the baseline]). We also included an unpublished GCM variant described in the preregistration for Study 1. However, we dropped it from the analyses because it did not account well for the data. We do not discuss it further as the theoretical grounds on which it was based are beyond the scope of this paper.

On the basis of these model assumptions, we predicted the category probabilities for all possible transfer items and selected an approximately equal number of items from the predefined categories. More specifically, we selected potential transfer items that strongly differed on their predictions between all models. On the basis of the predictions of these transfer items, we then simulated the responses of 100 participants for each (data-generating) model, by sampling individual responses from a binomial distribution. We further included variations in the number of transfer responses for each item between the simulation runs in order to find the optimal (ecological) design. After this, each model was fit via maximum likelihood on the simulated transfer choices for each participant individually, free fit in Study 3, or by using a cross-validation technique (i.e., leave-one-out) in Study 1.

We repeated these samplings and finally selected a design in each study that provided at least 80%-correct model recovery based on model Bayesian Information Criterion (BIC; Schwarz, 1978) values (e.g., such that the predictions of the baseline or

memory GCM versions could not be well mimicked by a CAM with freely varying feature weights). In Study 2, which we conducted last, we only considered the theoretical results from Study 1 in a Bayesian prediction method, which focuses on the similarity hypothesis. That is, we used the GCM parameter posteriors from Study 1 to make transfer predictions (all items) based on potential training items for Study 2. We then identified which items and manipulations were most sensitive to changes in similarity gradients despite including a large degree of parameter uncertainty in the simulation.

Table A1

Novel Transfer Stimuli in Study 1

Food item	Dimension		
	1()	2 ()	3()
1	2	3	1
2	1	4	1
3	2	4	1
4	3	4	1
5	2	1	3
6	1	1	4
7	2	1	4
8	3	1	4
9	1	3	4
10	2	2	3
11	1	4	3
12	4	1	1
13	4	3	2
14	1	1	2
15	3	4	2

Note. The dimensions correspond to those described for the training items (same order).

Table A2

Novel Transfer Stimuli in Study 2

Food item	Dimension		
	1()	2 ()	3()
1	4	5	1
2	5	4	1
3	5	5	1
4	4	4	2
5	3	4	1
6	2	4	1
7	5	2	1
8	1	4	1
9	2	3	5
10	4	2	4
11	3	3	5
12	5	2	3
13	3	3	2
14	2	4	2
15	1	4	3
16	1	5	2
17	2	5	2
18	5	1	3

Note. The dimensions correspond to those described for the training items (same order).

Table A3

Novel Transfer Stimuli in Study 3

Food item	Dimension		
	1()	2 ()	3()
1	1	2	2
2	2	1	2
3	5	1	1
4	2	1	3
5	1	1	4
6	5	5	1
7	2	5	3
8	5	1	3
9	3	5	4
10	4	5	4

Note. The dimensions correspond to those described for the training items (same order).

Table A4

Novel Reward-Judgment Items in Study 3

Food item	Dimension		
	1()	2 ()	3()
1	1	5	4
2	1	2	4
3	3	5	3
4	4	4	3
5	5	5	3
6	5	4	4
7	5	2	3
8	5	5	4
9	1	3	4
10	1	5	3
11	1	2	3
12	1	1	2
13	1	5	5
14	5	3	3
15	1	1	3
16	1	3	3
17	3	5	5
18	1	2	5
19	4	5	5
20	3	3	3
21	3	2	1
22	1	3	5
23	1	2	1
24	1	1	1
25	3	1	1

Note. The dimensions correspond to those described for the training items (same order).

Appendix B

Participant Exclusions and Modeling Contaminant Classification

Because our hypotheses were focused on the question of how reward affects category learning, we wanted to ensure that participants, indeed, learned the categorization task, especially in the online studies. Eventually, we decided to implement a Bayesian contaminant classification as introduced by Zeigenfuse and Lee (2011), assigning participants to guessing, medium-performance, and high-performance groups within each study and condition. The group means were nonhierarchically estimated with $\pi_{\text{guess}} = .5 < \pi_{\text{medium}} < \pi_{\text{high}}$, sampling $\pi_{\text{medium, high}} \sim \text{Uniform}(.5,1)$, and

individuals (k) were assigned to each group via $z_k \sim \text{Categorical}(\zeta)$, sampled from a Dirichlet distribution estimating the latent mixture of these groups, with a prior of $\zeta \sim \text{Dirichlet}(1, 1, 1)$. The accuracy estimate for the selected group then served as the latent probability of the number of correct responses for the participant in the last two training blocks, assuming a binomial response distribution with the corresponding number of observations. We excluded all participants who were assigned to the guessing model in more than 90% of the samples (the models converged on three chains with 20,000 iterations).

Because the performance was very low in Study 2, only one participant was confidently assigned to the guessing group, although a number of participants performed below chance. Instead of using a lower confidence criterion (i.e., 80%, as preregistered for Study 2), we decided to increase the number of observations and used the last three training blocks. However, alternatively using the lower confidence criterion (as preregistered) had the same effect. We further considered other exclusion methods (e.g., extreme behavior, extremely low response times [as preregistered for Study 1], or a simple 50%-accuracy criterion at the end of the training), which slightly differed in their consequences regarding the mixed-model pairwise (significance) tests in Study 3 (also with respect to model convergence); these are reported in the Online Supplemental Material (Categorization Accuracy). This was not the case for Studies 1 and 2. For Study 3 this might seem crucial; however, as it was our first study, we used conservative (undirected) significance criteria (despite having directed hypotheses) and strict p-value adjustments, which affected the results more heavily than the exclusions. That is, when we used the same significance criteria as for Studies 1 and 2, the results of Study 3 were identical to those reported in the main article, even without participant exclusions.

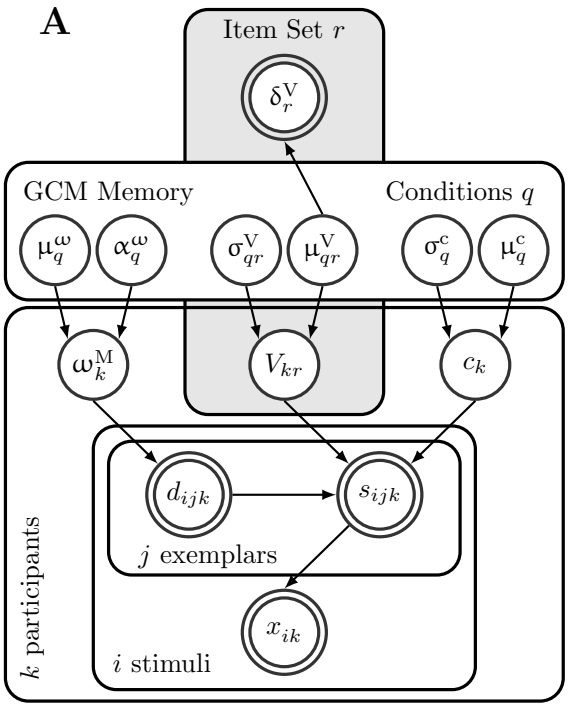
Appendix C

Model Descriptions and Priors

Memory GCM. For the memory GCM, we sampled the values of μ_{qr}^{V} from a Cauchy distribution with a mean of zero, passed to a lognormal distribution on the participant level to sample V_j (i.e., the distribution is normally distributed on the log scale). This method follows a reformalization of the GCM similarity function, that is, with $V_j e^{-cd} = e^{-cd+\ln(V_j)}$. Thus, the memory-strength estimate can be conveniently interpreted in terms of a similarity intercept/bias, such that, for example, a value of 0 in the distribution of $\mu_q^V r$ effectively becomes $V_j = e^0 = 1$ when sampled on the individual level (i.e., no different from low-reward items fixed to $V_j = 1$), and higher and lower estimates are symmetrically scaled (e.g., twice the memory strength: $2 = e^{.69}$, half the strength: $.5 = e^{-.69}$). We sampled one value for each participant in the memory GCM and applied it to all items of a set.

Similarity GCM. We sampled the gradient hyperparameters, μ_{qr}^{c} and σ_{qr}^{c} from truncated Cauchy and Gaussian distributions, respectively. The choice of the Gaussian prior for the standard deviation did affect the results (i.e., the precision constraint on this prior), which should be viewed with some caution (see Lee & Wagenmakers, 2014, p. 112). However, this choice was based on two important aspects. First, the prior is informed by the formal definition (or flexibility) of the GCM's generalization function. In short, there is hardly a difference in the predictions of a GCM with large values of $c_k j$ (~ 15) and one with extremely large values of the gradient (~ 100). Thus, sampling individual gradients from a population with a very tolerant (large) standard deviation will not contribute to parameter convergence, because medium and large individual values are virtually equally likely, although low similarity gradients are still covered. This leads to less certain μ_{qr}^{c} posteriors (wider CIs), as large standard deviations also allow a variety of means.

Second, and partly as a consequence of the first point, attempts to model the standard deviation (e.g., with more liberal sampling at the tails of the distributions using uniform or Cauchy distributions) led to frequent samples at the ceiling of the defined truncation range, as well as artificially skewed parameter posteriors. Thus, the model tended to prefer large standard deviations. Optimally, one would use fewer restrictions on the range of values to allow the parameter to converge at the most likely



GCM Memory

Attention & Exemplar Distance

$$\mu_q^{\omega} \sim \text{Dirichlet}(.25, .35, .4)$$
$$\alpha_q^{\omega} \sim \text{Cauchy}_{[.01,100]}(4.9, 1.4)$$
$$\omega_k \sim \text{Dirichlet}(\mu_q^{\omega}\alpha_q^{\omega}),\ k \text{ in } q$$
$$d_{ijk} \leftarrow \textstyle\sum_m \left(|p_i im - p_{jm}|\omega_{k(m)}\right)$$

Similarity Gradient

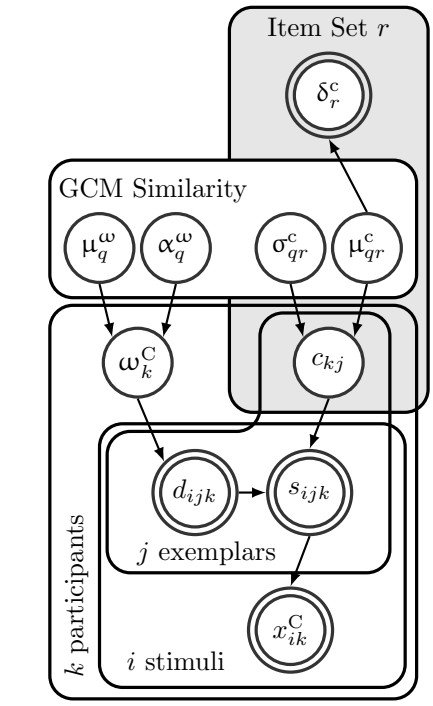
$$\mu_q^{\mathrm{c}} \sim \text{Cauchy}_{[0,30]}(2, 1)$$
$$\sigma_q^{\mathrm{c}} \sim \text{Gaussian}_{[0,15]}(3.7, 3.3)$$
$$c_k \sim \text{Gaussian}_{[0,30]}(\mu_q^{\mathrm{c}}, 1/(\sigma_q^{\mathrm{c}})^2),\ k \in q$$

Exemplar Memory Strength

$$\mu_{qr}^{\mathrm{V}} \sim \text{Cauchy}_{[-8,8]}(0, 1)$$
$$\sigma_{qr}^{\mathrm{V}} \sim \text{Gaussian}_{[0,15]}(3, .5)$$
$$\delta_r^{\mathrm{V}} \leftarrow \mu_{r[\text{Baseline}]}^{V} - \mu_{r[\text{Reward Condition}]}^{\mathrm{V}},\ \text{if } r = \text{high}$$
$$V_{kr} \begin{cases} \leftarrow 1; Set = 2[Study1\&2] & Set = 0[Study3] \\ \sim \text{Lognormal}(\mu_{qr}^{\mathrm{V}}, (\sigma_{qr}^{\mathrm{V}})^2) & r = high \end{cases}$$

Exemplar Similarities & Choice

$$s_{ijk} \leftarrow V_{kr}\mathrm{e}^{-c_{kj}d_{ijk}},\ k \in q,\ j \in r$$
$$x_{ik}^{\mathrm{M}} \leftarrow \frac{\sum_{j\in\text{Tami}} s_{ijk}}{\sum_{j\in\text{Tami}} s_{ijk} + \sum_{j\in\text{Humi}} s_{ijk}}$$



GCM Similarity

Attention & Exemplar Distance

$$\mu_q^{\omega} \sim \text{Dirichlet}(.25, .35, .4)$$
$$\alpha_q^{\omega} \sim \text{Cauchy}_{[.01,100]}(4.9, 1.4)$$
$$\omega_k \sim \text{Dirichlet}(\mu_q^{\omega}\alpha_q^{\omega}),\ k \in q$$
$$d_{ijk} \leftarrow \textstyle\sum_m \left(|p_i im - p_{jm}|\omega_{k(m)}\right)$$

Similarity Gradient

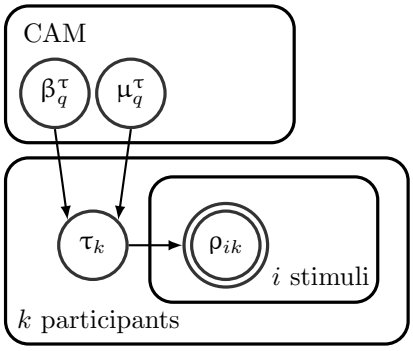
$$\mu_{qr}^{\mathrm{c}} \sim \text{Cauchy}_{[0,30]}(2, 1)$$
$$\sigma_{qr}^{\mathrm{c}} \sim \text{Gaussian}_{[0,15]}(3.7, 3.3)$$
$$\delta_r^{\mathrm{c}} \leftarrow \mu_{r[\text{Baseline}]}^{c} - \mu_{r[\text{Reward Condition}]}^{c}$$
$$c_k \sim \text{Gaussian}_{[0,30]}(\mu_{qr}^{\mathrm{c}}, 1/(\sigma_{qr}^{\mathrm{c}})^2),\ j \in r$$

Exemplar Similarities & Choice

$$s_{ijk} \leftarrow \mathrm{e}^{-c_{kj}d_{ijk}},\ k \in q,\ j \in r$$
$$x_{ik}^{\mathrm{C}} \leftarrow \frac{\sum_{j\in\text{Tami}} s_{ijk}}{\sum_{j\in\text{Tami}} s_{ijk} + \sum_{j\in\text{Humi}} s_{ijk}}$$

Figure C1. Graphical Bayesian model descriptions. (A) Generalized context model (GCM) with set-specific memory strengths. (B) GCM with set-specific generalization gradients. Shaded boxes in A and B highlight the parameter tests of interest, that is, differences in memory strength δ_r^V and similarity δ_r^c, on each item set (r), between conditions (see also Figure C2).

A



CAM

Dimension Weights (Importance)

$$\mu_{qm}^\tau \sim \begin{cases} \text{Cauchy}_{[-10,10]}(.8,5) & m=1 \\ \text{Cauchy}_{[-10,10]}(.7,5) & m=2 \\ \text{Cauchy}_{[-10,10]}(1.4,5) & m=3 \end{cases}$$

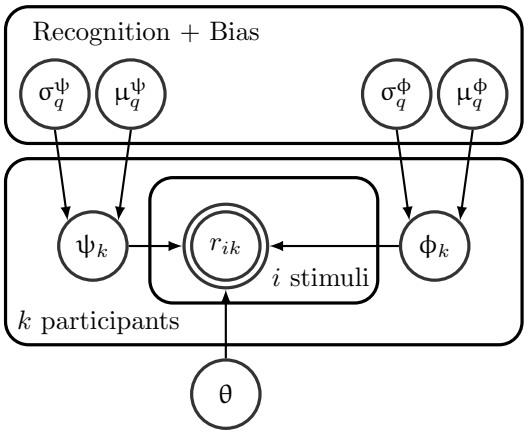
$$\sigma_{qm}^\tau \sim \begin{cases} \text{Cauchy}_{[0,5]}(1.5,4) & m=1 \\ \text{Cauchy}_{[0,5]}(1.5,5) & m=2 \\ \text{Cauchy}_{[0,5]}(.7,5) & m=3 \end{cases}$$

$$\tau_{qm} \sim \text{Gaussian}_{[-10,10]}(\mu_{qm}^\tau, 1/(\sigma_{qm}^\tau)^2)$$

Choice

$$\rho_{ik} \leftarrow \left(1+\mathrm{e}^{-\sum_m p_{im}\tau_{km}}\right)^{-1}$$

B

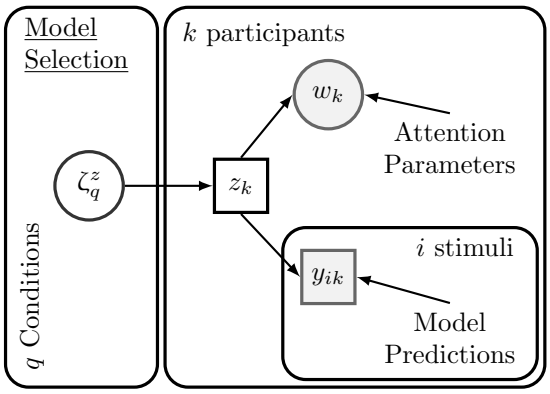


Recognition ψ & Bias ϕ

$$\mu_q^\psi \sim \text{Cauchy}_{[0,1]}(.5,1)$$
$$\sigma_q^\psi \sim \text{Cauchy}_{[0,1]}(0,1.5)$$
$$\mu_q^\phi \sim \text{Uniform}(0,.5)$$
$$\sigma_q^\phi \sim \text{Cauchy}_{[0,.5]}(0,1.5)$$
$$\mu_k^\psi \sim \text{Cauchy}(\mu_q^\psi, 1/(\sigma_q^\psi)^2)$$
$$\mu_k^\phi \sim \text{Cauchy}(\mu_q^\phi, 1/(\sigma_q^\phi)^2)$$
$$\theta \sim \text{Gaussian}_{[.01,1]}(1,4)$$

$$r_{ik} \leftarrow \begin{cases} 1-\psi_k & i \in \text{Trained(Tami)} \\ \psi_k & i \in \text{Trained(Humi)} \\ \theta\phi_k & i \in \text{Novel} \end{cases}$$

C



Model Selection

$$\zeta_q^z \sim \text{Dirichlet}(1,1.5,1.5,1.5)$$
$$z_k \sim \text{Categorical}(\zeta_q^z),\ k \in q$$

Categorization Behavior

$$y_{ik} \sim \begin{cases} \text{Binomial}(r_{ik}) & z=1 \\ \text{Binomial}(x_{ik}^{\mathrm{M}}) & z=2 \\ \text{Binomial}(x_{ik}^{\mathrm{C}}) & z=3 \\ \text{Binomial}(\rho_{ik}) & z=4 \end{cases}$$

Common Cause: Attention

$$w_k \sim \begin{cases} \text{Dirichlet}(1,1,1) & z=1 \\ \text{Dirichlet}(\omega_k^{\mathrm{M}}) & z=2 \\ \text{Dirichlet}(\omega_k^{\mathrm{C}}) & z=3 \\ \text{Dirichlet}(\tau_k) & z=4 \end{cases}$$

Figure C2. Graphical Bayesian model descriptions. (A) Cue-abstraction model (CAM). The stimulus values p_{im} given to CAM were mean-centered on their corresponding scale, reflecting an unbiased standard regression. (B) Recognition and bias model, and (C) Latent mixture model selection, and common cause modeling (see text).

values. However, as described, extreme values of the similarity gradient are difficult to interpret. Thus, we forced the hyperposterior distribution of the gradient standard deviation into a Gaussian shape, to increase the model's commitment to more focused estimates and, in turn, more specific population mean estimates. This seems in line with arguments by Lee and Vanpaemel (2018), as well as Rouder, Morey, Verhagen, Swagman, and Wagenmakers (2017; see Choice of Priors section). However, there are other ways of sampling the similarity gradient that we did not explore (e.g., a Gamma distribution; see Bartlema et al., 2014).

We sampled individual c_{kj} for each exemplar in set r from the hyperdistribution. That is, c_{kj} does not distinguish between the distribution of participants and of exemplars in an item set. More complex solutions to disentangle item and participant distributions seem statistically unreasonable in our design (e.g., implementing by-participant sampling distributions for item sets with two exemplars). Simpler models (e.g., sampling c_{kj} as done for V_{kr}) did not converge in all studies. Thus, we applied the converging solution to each study, which also seems theoretically consistent if one opens oneself to the idea that similarity gradients are not equal for all exemplars. However, please note, this solution makes the model more flexible than in its usual implementation.

Recognition and bias model. The parameter ψ estimates how accurate the transfer phase decisions are for trained items (as deviation from perfect recognition), regardless of the category. The parameter ϕ estimates a choice bias for novel transfer items. The model, thus, covers a range of behavior, from overall guessing to recognition to "unidentified strategies" for novel items. Because the latter may vary widely between participants, we expanded this term with the parameter ϵ (i.e., one value for the whole population), introducing random noise into the estimates, which also eliminated issues with auto-correlations on ϕ.

Choice of priors. Single model fits were used to define pseudopriors (data-driven priors) for two main reasons. First, Rouder et al. (2017, p. 309f) recommended fine-tuning precision priors to test for parameter differences via BFs, especially with unscaled (unstandardized) parameters. That is, if the prior precision is set too high or too low, it will include either too many unreasonable or too few reasonable parameter values, respectively. Fine-tuning the priors, thus, prevents a too

Table C1

Single Model Fits and Pseudopriors for Model Selection

Model priors	Study 1			Study 2			Study 3			
	Posteriors		Selection prior	Posteriors		Selection prior	Posteriors			Selection prior
	C	R		C	R		C	TR	AR	
CAM										
$\mu^{\beta 1} \sim \text{Cauchy}_{[-10,10]}(1, 0.1)$	1.4 (49.8)	1.2 (58.4)	1.3 (5.0)	0.1 (14.2)	0.7 (16.1)	0.5 (5.0)	0.9 (46.9)	0.8 (43.2)	0.7 (64.5)	0.8 (5.0)
$\mu^{\beta 2} \sim \text{Cauchy}_{[-10,10]}(1, 0.1)$	0.5 (95.0)	0.1 (132.2)	0.3 (5.0)	0.6 (20.9)	1.0 (23.4)	0.7 (5.0)	0.8 (41.8)	0.8 (37.6)	0.4 (58.8)	0.7 (5.0)
$\mu^{\beta 3} \sim \text{Cauchy}_{[-10,10]}(1, 0.1)$	0.4 (140.1)	0.3 (145.4)	0.3 (5.0)	1.4 (5.1)	1.1 (13.8)	1.2 (5.0)	1.5 (18.1)	1.5 (13.3)	1.2 (27.2)	1.4 (5.0)
$\sigma^{\beta 1} \sim \text{Cauchy}_{[0,5]}(0.1, 0.1)$	0.6 (87.4)	0.7 (66.3)	0.6 (5.0)	0.9 (8.8)	0.5 (39.7)	0.5 (5.0)	1.7 (4.5)	1.4 (8.3)	1.9 (4.4)	1.5 (4.0)
$\sigma^{\beta 2} \sim \text{Cauchy}_{[0,5]}(0.1, 0.1)$	1.4 (14.4)	2.0 (5.9)	1.7 (5.0)	1.0 (10.4)	0.7 (25.5)	0.5 (5.0)	1.5 (5.8)	1.2 (10.4)	1.7 (5.4)	1.5 (5.0)
$\sigma^{\beta 3} \sim \text{Cauchy}_{[0,5]}(0.1, 0.1)$	1.9 (8.2)	2.0 (7.0)	2.0 (5.0)	0.2 (284.1)	0.3 (98.1)	0.5 (5.0)	0.6 (34.7)	0.4 (86.5)	0.8 (25.2)	0.7 (5.0)
GCM										
$\mu^{c} \sim \text{Cauchy}_{[0,30]}(5, 0.5)$	2.07 (8.03)	2.16 (26.85)	2.0 (1.0)	0.37 (10.47)	0.66 (4.62)	2.0 (1.0)	2.94 (1.19)	2.58 (0.87)	1.28 (1.54)	2.0 (1.0)
$\sigma^{c} \sim \text{Gaussian}_{[0,15]}(1, 0.5)$	1.76 (11.49)	1.31 (29.54)	1.5 (5.0)	2.97 (8.65)	3.06 (8.12)	3.0 (5.0)	3.26 (3.21)	4.04 (2.71)	3.75 (3.86)	3.7 (3.3)
$\mu^{\omega 1} \sim$	0.61	0.63	0.6	0.3	0.3	0.3	0.24	0.27	0.27	0.25
$\mu^{\omega 2} \sim \text{Dirichlet}(1, 1, 1)$	0.2	0.17	0.2	0.34	0.3	0.3	0.39	0.38	0.30	0.35
$\mu^{\omega 3} \sim$	0.2	0.2	0.2	0.37	0.4	0.4	0.37	0.35	0.43	0.4
$\alpha^{\omega} \sim \text{Cauchy}_{[.01,100]}(3, 4)$	2.69 (13.52)	3.09 (11.65)	2.9 (5.0)	3.06 (10.75)	2.86 (14.24)	3.0 (5.0)	6.52 (0.48)	3.94 (2.14)	4.15 (1.68)	4.9 (1.4)

Note. Single model priors and posterior estimates. Numbers in parentheses indicate the precision of the corresponding distribution (i.e., $1/\sigma^2$). The distributions between conditions were roughly averaged to yield identical pseudopriors in every condition of a study for the model selection (Selection prior). Large precision values were reset to 5 to retain some flexibility in the mixture model. Please note, larger precision reflects that the range of likely parameter values under a model is more certain, given the data. The priors for the similarity gradients were, however, chosen to be of the same scale (i.e., 1) in every study, to allow us to directly compare the evidence (Bayes factors) between studies. The m dimensions in Study 2 were coded as "original" dimensions in Table 1 of the main manuscript. CAM = Cue-abstraction model; GCM = generalized context model; C and R refer to the control and reward conditions, respectively. TR = typical reward; AR = atypical reward.

high or too low prior mass for zero parameter differences (e.g., for the null hypothesis $\delta_r^c = \mu_{r[\text{Baseline}]}^c - \mu_{r[\text{Reward Condition}]}^c = 0$), which affects the conclusions based on the obtained BFs in parameter comparison. Second, with respect to latent class model selection, as performed in our studies, it has been discussed that transdimensional model selection methods can falsely favor unlikely models by sampling unlikely parameters for likely models (see Sisson, 2005), which can become more severe in large model and parameter spaces. Using pseudo (data-driven) priors is one possible solution to this potential issue. However, in our studies, using different (or less-informed) priors basically led to the same results.

For obtaining the pseudopriors, we used single model fits with relatively uninformative priors in a plausible range of parameter values (see Table C1). As described, the purpose of this method was tuning the priors with respect to the model selection and the parameter tests. To avoid biasing our hypothesis tests with the data-driven priors, the single-model GCM did not contain exemplar-specific parameters, that is, assuming one similarity gradient per participant, while setting the memory strength to 1 for all items (we also left out the common cause constraint on feature attention). As described, an advantage of obtaining the pseudopriors from the "null-effect" version of the GCM is a reasonable assumption about the prior scale (precision) having "one common" similarity-gradient parameter in the current task. However we used a more liberal prior scale for μ_{rq}^c (i.e., precision set to 1) than the actual single-model estimates to avoid an inflation of the resulting BFs. That is, to obtain prior likelihoods for the null hypothesis (i.e., for $\delta_r^c = 0$ and $\delta_r^V = 0$), we took two of the defined priors of the corresponding hyper means (e.g., μ_q^c) and simply subtracted them from each other. If their precision had been larger (e.g., 5), this would have substantially increased the reported BF in favor of the alternative, as this prior scale would reflect a very sceptical belief about the existence of an effect, and, thus, a stronger change toward a new (less sceptical) belief. To obtain the BFs, finally, we also calculated the parameter differences between the conditions (e.g., $\mu_{r[\text{baseline}]}^c - \mu_{r[\text{AR}]}^c$ in Study 3) and compared them to the prior via the SD density ratio (Dickey, 1971) in the interval of [-.01,.01]. We used the same prior scales for the similarity gradient and the memory strength for all studies and conditions, for convenience and consistency in interpreting the BFs.